

# ON THE ASYMPTOTIC THEORY FOR LEAST SQUARES SERIES: POINTWISE AND UNIFORM RESULTS

ALEXANDRE BELLONI, XIAOHONG CHEN, VICTOR CHERNOZHUKOV, AND KENGO KATO

**ABSTRACT.** In this work we consider series estimators for the conditional mean in light of three new ingredients: (i) sharp LLNs for matrices derived from the non-commutative Khinchin inequalities, (ii) bounds on the Lebesgue constant that controls the ratio between the  $L^\infty$  and  $L^2$ -norms, and (iii) maximal inequalities with data-dependent bounds for processes whose entropy integrals diverge at some rate.

These technical tools allow us to contribute to the series literature, specifically the seminal work of Newey (1995), as follows. First, we weaken considerably the condition on the number  $k$  of approximating functions used in series estimation from the typical  $k^2/n \rightarrow 0$  to  $k/n \rightarrow 0$ , up to log factors, which was available only for splines before. Second, under the same weak conditions we derive  $L^2$  rates and pointwise central limit theorems results when the approximation error vanishes. Under a incorrectly specified model, i.e. when the approximation error does not vanish, analogous results are also shown. Third, under stronger conditions we derive uniform rates and functional central limit theorems that holds if the approximation error vanishes or not. That is, we derive the strong approximation for the entire estimate of the non-parametric function. Finally, we derive uniform rates and inference results for linear functionals of interest of the conditional expectation function such as its partial derivative or conditional average partial derivative.

## 1. INTRODUCTION

Series estimators have been playing a central role on various fields. In econometric applications it is common that the exact form of a conditional expectation is unknown and having a flexible functional form can lead to improvements over a pre-specified functional form. Series estimation offers exactly that by approximating the unknown function based on  $k$  basic functions, where  $k$  is allowed to grow with the sample size  $n$  to balance the trade off between variance and bias.

Several asymptotic properties of series estimators have been investigated in the literature. The focus has been on convergence rates and asymptotic normality results (see

---

*Date:* First version: May 2006, This VERSION OF DECEMBER 4, 2012.

Andrews, 1991; Eastwood and Gallant, 1991; Gallant and Souza, 1991; Newey, 1997, and the references therein).

This work revisits the topic by making use of three critical ingredients:

- i. The sharp LLNs for matrices derived from the non-commutative Khinchin inequalities.
- ii. The sharp bounds on the Lebesgue constant that controls the ratio between the  $L^\infty$  and  $L^2$ -norms of the least squares approximation of functions (which is bounded or grows like a  $\log k$  in many cases).
- iii. Maximal inequalities with data-dependent bounds for processes whose entropy integrals diverge at some rate.

To the best of our knowledge, these results are the first applications of the first ingredient to statistical estimation problems. Regarding the second ingredient, it has already been used by Huang (2003a) but for splines only. The third ingredient was derived to allow for weak moment conditions. All of these ingredients are critical for generating sharp results.

This approach allows to contribute to the series literature in several directions. First, we weaken considerably the condition on the number  $k$  of approximating functions used in series estimation from the typical  $k^2/n \rightarrow 0$  (see Newey, 1997) to

$$k/n \rightarrow 0 \text{ (up to logs)}$$

for bounded basis which was available only for splines before (Huang, 2003a; Stone, 1994). Second, under the same weak conditions we derive  $L^2$  rates and pointwise central limit theorems results when the approximation error vanishes. Under a misspecified model, i.e. when the approximation error does not vanish, analogous results are also shown. Third, under stronger conditions we derive uniform rates and functional central limit theorems that hold if the approximation error vanishes or not. By the functional central limit theorem we mean here that the entire estimate of the non-parametric function is uniformly close to a Gaussian process that can change with  $n$ . That is, we derive the strong approximation for the entire estimate of the non-parametric function.

Another set of results established here pertains to the estimation and inference methods for linear functionals  $\theta$  of the conditional mean function  $g : \mathcal{X} \rightarrow \mathbb{R}$ . Examples of linear functionals  $\theta$  of interest include, for  $x = (w, v) \in \mathcal{X}$  and  $x_j$  denoting the  $j$ -th component of  $x$ ,

1. the partial derivative:  $\theta(x) = \partial_{x_j} g(x)$ ;

2. the average partial derivative:  $\theta = \int \partial_{x_j} g(x) d\mu(x)$ ;
3. the conditional average partial derivative:  $\theta(w) = \int \partial_{x_j} g(w, v) d\mu(v|w)$ .

where the measure  $\mu$  entering the definitions above are taken as known; the result can be extended to include estimated measures. Under weak conditions we derive pointwise results for rates of convergence, large sample distributions and inference methods based on the Gaussian approximation. Under stronger conditions we derive new strong approximation for the entire estimate of the non-parametric function. Specifically, we derive uniform results for rates of convergence, large sample distributions and inference methods based on the Gaussian approximation.

**Notation.** In what follows, all parameter values are indexed by the sample size  $n$ , but we omit the index whenever this does not cause confusion. We use the notation  $(a)_+ = \max\{a, 0\}$ ,  $a \vee b = \max\{a, b\}$  and  $a \wedge b = \min\{a, b\}$ . The  $\ell_2$ -norm of a vector  $v$  is denoted by  $\|v\|$ , while for a matrix  $Q$  the maximum eigenvalue is denoted by  $\|Q\|$ . We also use standard notation in the empirical process literature,

$$\mathbb{E}_n[f] = \mathbb{E}_n[f(w_i)] = \sum_{i=1}^n f(w_i)/n,$$

and we use the notation  $a \lesssim b$  to denote  $a \leq cb$  for some constant  $c > 0$  that does not depend on  $n$ ; and  $a \lesssim_P b$  to denote  $a = O_P(b)$ . Moreover, for two random variables  $X, Y$  we say that  $X =_d Y$  if they have the same probability distribution.

## 2. SET-UP

Consider a sequence of models, indexed by the sample size  $n$ ,

$$y_i = g(x_i) + \epsilon_i, \quad E[\epsilon_i | x_i] = 0, \quad x_i \in \mathcal{X} \subseteq \mathbb{R}^d, \quad i = 1, \dots, n, \quad (2.1)$$

where  $y_i$  is the response variable and  $x \mapsto g(x) = E[y_i | x_i = x] \in \mathcal{G}_n$  a class of functions. We assume that  $\mathcal{X}$  is a compact set in  $\mathbb{R}^d$  and that all conditions stated below hold uniformly in  $n$ . For notational convenience we omit indexing by  $n$  in what follows.

**Assumption A.1** *Random vectors  $(\epsilon_i, x_i)'$ ,  $i = 1, \dots, n$ , are i.i.d. and  $\sup_{i \leq n} \sigma_i^2 = E[\epsilon_i^2 | x_i]$  is bounded.*

We approximate the function  $x \mapsto g(x)$  by linear forms  $x \mapsto p(x)'b$ , where

$$x \mapsto p(x) = (p_j(x), j = 1, \dots, k)$$

is a vector of approximating functions that can change with  $n$ . We denote the regressors as

$$p_i = p(x_i) = (p_j(x_i), j = 1, \dots, k)$$

where we use  $i$  to index observations  $p(x_i)$ , and  $j$  to index components of  $p(x_i)$ . The next assumption impose regularity conditions on the regressors.

**Assumption A.2** *Eigenvalues of  $Q := E[p_i p_i']$  are bounded above and away from zero. Also we let*

$$\xi_k := \sup_{x \in \mathcal{X}} \|Q^{-1/2} p(x)\|,$$

and that  $k$  is chosen so that

$$\xi_k^2 \log n / n \rightarrow 0. \quad (2.2)$$

**Normalization.** *To simplify notation, we normalize  $Q = I$ , but we shall treat  $Q$  as unknown, that is we deal with random design.*

The relation (2.2) restricts how fast  $k$  can grow with  $n$  but it is a mild condition for many interesting basis of functions as discussed below. Condition A.2 also imposes that  $p_i$ 's are not too co-linear. The following proposition establishes a simple sufficient condition for A.2 based on orthonormal bases with respect to a different measure.

**Proposition 1** (Stability of Bounds on Singular Values). *Let  $x \sim F$  and the regressors  $p_i = p(x_i)$ , with  $x \mapsto p(x)$  orthonormal on  $(\mathcal{X}, \mu)$  for some measure  $\mu$ . Then A.2 is satisfied if  $dF/d\mu$  is bounded above and away from zero.*

It is well known that the least squares parameter solves

$$\beta = \arg \min_{b \in \mathbb{R}^k} E[(y_i - p(x_i)'b)^2],$$

which by (2.1) implies that  $\beta$  also solves

$$\beta = \arg \min_{b \in \mathbb{R}^k} E[(g(x_i) - p(x_i)'b)^2].$$

We call  $x \mapsto g(x)$  the target function and  $x \mapsto g_k(x) = p(x)' \beta$  the surrogate function. In this setting, the surrogate function provides the best linear approximation to the target function.

Accordingly we have a many regressors model

$$y_i = p_i' \beta + u_i, \quad E[u_i x_i] = 0, \quad u_i := r_i + \epsilon_i$$

where

$$r_i := r(x_i) = g(x_i) - p(x_i)' \beta,$$

is the approximation error. The least squares estimator of  $\beta$  is

$$\hat{\beta} = \arg \min_{b \in \mathbb{R}^k} \mathbb{E}_n[(y_i - p_i' b)^2],$$

which induces the estimator  $\hat{g}(x) := p(x)' \hat{\beta}$  for the target function  $g(x)$ . Thus, it follows that we can decompose the error in estimating the target function as

$$\hat{g}(x) - g(x) = p(x)'(\hat{\beta} - \beta) + r(x),$$

where the first term in the right hand side is the estimation error and the second term is the approximation error.

We are also interested in various quantities  $\theta$  created as linear functionals of the conditional mean function. As discussed in the introduction, examples include the partial derivative function, the average partial derivative function, and the conditional average partial derivative. By the linearity of the series approximations, the above parameters can be seen as linear functions of the least squares coefficients  $\beta$  up to an approximation error. Importantly, in each example above we could be interested in estimating  $\theta(w)$  simultaneously for many values of  $w \in \mathcal{W}$ . We let  $\mathcal{I} \subset \mathcal{W}$  denote the set of indices of interest. By the linearity of the series approximations, the above parameters can be seen as linear functions of the least squares coefficients  $\beta$  up to an approximation error, that is

$$\theta(w) = \ell(w)' \beta + r_n(w), \quad w \in \mathcal{I}, \quad (2.3)$$

where  $\ell(w)' \beta$  is the series approximation, with  $\ell(w)$  denoting the  $k$ -vector of loadings on the coefficients, and  $r_n(w)$  is the remainder term, which corresponds to the approximation error. Indeed, the decomposition (2.3) arises from the application of different linear operators  $\mathcal{A}$  to the decomposition  $g(\cdot) = p(\cdot)' \beta + r(\cdot)$  and evaluating the resulting functions at  $w$ :

$$(\mathcal{A}g(\cdot)) [w] = (\mathcal{A}p(\cdot)) [w]' \beta + (\mathcal{A}r(\cdot)) [w]. \quad (2.4)$$

Examples of the operator  $\mathcal{A}$  corresponding to the cases enumerated in the introduction are given by, respectively,

1. a differential operator:  $(\mathcal{A}f)[x] = (\partial_{x_j} f)[x]$ , so that

$$\ell(x) = \partial_{x_j} p(x), \quad r_n(x) = \partial_{x_j} r(x) ;$$

2. an integro-differential operator:  $\mathcal{A}f = \int \partial_{x_j} f(x) d\mu(x)$ , so that

$$\ell = \int \partial_{x_j} p(x) d\mu(x), \quad r_n = \int \partial_{x_j} r(x) d\mu(x) ;$$

3. a partial integro-differential operator:  $(\mathcal{A}f)[w] = \int \partial_{x_j} f(x) d\mu(v|w)$ , so that

$$\ell(w) = \int \partial_{x_j} Z(x) d\mu(v|w), \quad r_n(w) = \int \partial_{x_j} r(x) d\mu(v|w).$$

For notational convenience, we use the formulation (2.3) in the analysis, instead of the motivational formulation (2.4).

We shall provide the inference tools that will be valid for inference on the series approximation

$$\ell(w)' \beta, \quad w \in \mathcal{I},$$

and, provided that the approximation error  $r_n(w)$ ,  $w \in \mathcal{I}$ , is small enough as compared to the estimation noise, these tools will also be valid for inference on the functional of interest:

$$\theta(w), \quad w \in \mathcal{I}.$$

In this case, the series approximation is an important intermediary target, whereas the functional  $\theta$  is the ultimate target. The inference will be based on the plug-in estimator  $\hat{\theta}(w) := \ell(w)' \hat{\beta}$  of the the series approximation  $\ell(w)' \beta$  and hence of the final target  $\theta(w)$ .

### 3. APPROXIMATION PROPERTIES OF LEAST SQUARES

Next we consider approximation properties of the least squares estimator. Not surprisingly, approximation properties must rely on the particular choice of approximating functions. At this point it is instructive to consider particular examples of relevant basis used in the literature.

**Example 1** (Polynomial series). Consider  $\mathcal{X} = [0, 1]$  and polynomial series given by

$$p(x) = (1, x, x^2, \dots, x^{k-1}).$$

In order to reduce collinearity problems, orthonormalize with respect to the uniform measure to get the Legendre polynomials

$$p(x) = (1, x, 2^{-1}(3x^2 - 1), \dots)$$

with

$$\xi_k \lesssim k.$$

**Example 2** (Fourier series). Consider the domain  $\mathcal{X} = [0, 1]$  and a fourier series given by

$$p(x) = (1, \cos(2\pi j), \sin(2\pi j), j = 1, 2, \dots, k/2 - 1),$$

for  $k$  even, are orthonormal with respect to the Lebesgue measure, with

$$\xi_k \lesssim \sqrt{k}.$$

**Example 3** (Splines). Let  $\mathcal{X} = [-1, 1]$  and consider the linear spline series, or spline series of order 1, with a finite number of equally spaced knots  $k_1, k_2, \dots, k_r$  in  $\mathcal{X}$ :

$$p(x) = (1, x, (x - k_1)_+, \dots, (x - k_r)_+)'.$$

The cubic spline series takes the form:

$$p(x) = (1, (x, x^2, x^3), (x - k_1)_+^3, \dots, (x - k_r)_+^3)'.$$

The function  $x \mapsto p(x)'b$  constructed using cubic splines is twice differentiable in  $x$  for any  $b$ . Instead of pure splines, we often use B-splines, which are linear transformations of the above functions with lower multicellularity; moreover,

$$\xi_k \lesssim \sqrt{k}.$$

**Example 4** (Cohen-Deubechies-Vial wavelet bases). Let  $\mathcal{X} = [0, 1]$  and consider Cohen-Deubechies-Vial (CDV) wavelet bases. See Section 4 in Cohen et al. (1993) and Chapter 7.5 in Mallat (2009) for details on CDV wavelet bases. CDV wavelet bases are a class of orthonormal bases of  $L^2[0, 1]$ , which is the standard  $L^2$  space for functions defined on  $[0, 1]$ . Each such basis is built from a Deubechies scaling function  $\phi$  (defined on  $\mathbb{R}$ ) and the wavelet  $\psi$  of order  $N$  starting from a fixed resolution level  $J_0$  such that  $2^{J_0} \geq 2N$ . The  $\phi$  and  $\psi$  are supported on  $[0, 2N - 1]$  and  $[-N + 1, N]$ , respectively. Translate  $\phi$  so that  $\phi$  has support  $[-N + 1, N]$ . Let

$$\phi_{lm} = \phi(2^l \cdot -m), \quad \psi_{lm} = \psi(2^l \cdot -m), \quad l, m \geq 0.$$

The  $\phi_{J_0 m}, \psi_{lm}$  that are supported in the interior of  $[0, 1]$  are all kept ( $\phi_{J_0 m}^{\text{int}} = \phi_{J_0 m}$  for  $m = N, \dots, 2^{J_0} - N - 1$ ;  $\psi_{lm}^{\text{int}} = \psi_{lm}$  for  $m = N, \dots, 2^l - N - 1, l \geq J_0$ ), and suitable boundary corrected functions are added, so that  $\{\phi_{J_0 m}^{\text{int}}\}_{m=0}^{2^{J_0}-1} \cup \{\psi_{lm}^{\text{int}}\}_{0 \leq m < 2^l, l \geq J_0}$  forms an orthonormal basis of  $L^2[0, 1]$ . Suppose that  $k = 2^J$  for some  $J > J_0$ . Let

$$p(x) = (\phi_{J_0, 0}(x), \dots, \phi_{J_0, 2^{J_0}-1}(x), \psi_{J_0, 0}(x), \dots, \psi_{J-1, 2^{J-1}-1}(x))'.$$

Then

$$\xi_k \lesssim \sqrt{k}.$$

CDV wavelet bases are useful for approximating not necessarily periodic functions.

**Example 5** (Tensor Products). Generalizations to multiple regressors are straightforward using tensor products of unidimensional series. Suppose that the basic regressors are

$$x_i = (x_{1i}, \dots, x_{di}),$$

then we can create  $d$  series for each basic regressor, then create all interactions of the  $d$  series, called tensor products, and collect them into regressor vector  $p_i$ . If each series for a basic regressor has  $J$  terms, then the final regressor has dimension

$$k \lesssim J^d,$$

which explodes exponentially in the dimension  $d$ . The bounds on  $\xi_k$  in terms of  $k$  remain the same as in one-dimensional case.

Each base described in Examples 1-5 has different approximation properties which also depend on the particular class of functions  $\mathcal{G}_n$ . The following captures the essence of this dependence into two quantities.

**Assumption A.3** *For each  $g \in \mathcal{G}_n$  and integer  $k \geq 1$ , there are finite constants  $c_k$  and  $\ell_k$  such that*

$$\|r\|_{F,2} := \sqrt{\int r^2(x) dF(x)} \leq c_k \quad \text{and} \quad \|r\|_{F,\infty} := \sup_{x \in \mathcal{X}} |r(x)| \leq \ell_k c_k,$$

where we call  $\ell_k$  the generalized Lebesgue constant.

These quantities characterize the approximation properties of the underlying class of functions under  $L_2$  and uniform distances. Next we discuss primitive bounds on them.

**3.1. Bounds on  $c_k$ .** In what follows we call the case where  $c_k \rightarrow 0$  the correctly specified case. In particular, if for every  $n$  large enough the series are formed from bases that span  $\mathcal{G}_n$ , then  $c_k \rightarrow 0$  as  $k \rightarrow \infty$ . However, if series are formed from bases that do not span  $\mathcal{G}_n$ , then  $c_k \rightarrow c_\infty$  as  $k \rightarrow \infty$  where potentially  $c_\infty > 0$ . We call any case where  $c_k \not\rightarrow 0$  the incorrectly specified case.

Moreover, since

$$\inf_b \|g - p'b\|_{F,2} \leq c_k \leq \inf_b \|g - p'b\|_{F,\infty},$$

the approximation rates  $c_k$  are readily available from rates  $c_k$  computed in Approximation Theory (see DeVore and Lorentz (1993)). For example, if  $\mathcal{G}_n$  is  $s$ -smooth, namely a Holder class of smoothness order  $s$ , then

$$c_k \lesssim k^{-s/d}$$



for the examples of series given above, and

$$c_k \lesssim k^{-[s \wedge s_0]/d}$$

for splines of order  $s_0$ . However, we do not have to specify  $\mathcal{G}_n$  in terms of smoothness.

**3.2. Bounds on  $\ell_k$ .** A least squares approximation by a particular series for the function class  $\mathcal{G}_n$  is called co-minimal if the generalized Lebesgue constant  $\ell_k$  is small in the sense of being a slowly varying function in  $k$ .

A valid (arguably crude) bound on  $\ell_k$ , which is independent of  $\mathcal{G}_n$ , is

$$\ell_k \leq \xi_k + 1,$$

which is not small since  $\xi_k \gtrsim \sqrt{k}$  for many interesting basis. Much sharper bounds follow from Approximation Theory for some important cases. We list a few examples next.

**Example 6** (Fourier series, continued). For Fourier series on  $\mathcal{X} = [0, 1]$ ,  $F = U(0, 1)$ , and  $\mathcal{G}_n \subset C(\mathcal{X})$

$$\ell_k \leq C_0 \log k + C_1,$$

where here and below  $C_0$  and  $C_1$  are some universal constants.

**Example 7** (B-splines, continued). For B-splines of order  $s$  on  $\mathcal{X} = [0, 1]$ ,  $F = U(0, 1)$ , and  $\mathcal{G}_n \subset C(\mathcal{X})$

$$\ell_k \leq C_0,$$

under approximately uniform placement of knots.

**Example 8** (Chebyshev polynomials). For Chebyshev polynomials on  $\mathcal{X} = [-1, 1]$ ,  $dF(x)/dx = 1/\sqrt{1-x^2}$ , and  $\mathcal{G}_n \subset C(\mathcal{X})$

$$\ell_k \leq C_0 \log k + C_1.$$

**Example 9** (Local polynomials). For local polynomials of order  $s$  on  $\mathcal{X} = [0, 1]$ ,  $F = U(0, 1)$ , and  $\mathcal{G}_n \subset G$ , a Holder class,

$$\ell_k \leq C_0.$$

**Example 10** (Tailored Function Classes). For each type of series approximations, it is possible to specify function classes for which the generalized Lebesgue constants are small.

Since the Lebesgue constant depends on the particular basis and on the underlying probability measure, it is important to have a stability result for the Lebesgue constant. The next proposition provides a bound on  $\ell_k c_k$  for most functions in the  $\alpha$ -ellipsoid class

$$\mathcal{F}(\alpha) = \left\{ \sum_{j \geq 1} p_j(x) j^{-\alpha} \xi_j : \xi_j \in \mathbb{R}, j \geq 1 \right\}$$

according to a Gaussian measure on the coefficients  $\xi_j$ ,  $j \geq 1$ , provided the basis functions are bounded and Lipschitz.

**Proposition 2** (Generic Stability of Approximation Error for  $\alpha$ -Ellipsoid). *Consider the standard Gaussian measure on the coefficients  $\xi_j$ ,  $j \geq 1$ , let  $f = \sum_{j \geq 1} p_j(x) j^{-\alpha} \xi_j$  and let  $\ell_k(f)$  and  $c_k(f)$  denote respective the generalized Lebesgue constant and the  $L_2$  approximation rate associated with  $f$ . If the basis  $\{p_j(x)\}_{j \geq 1}$  obey  $\sup_{x \in \mathcal{X}, j \geq 1} |p_j(x)| \lesssim 1$  and  $\sup_{x \in \mathcal{X}} \|\nabla p_j(x)\| \leq M_j$ , with  $j^{-\alpha} M_j \log^{1/2} j = o(1)$  as  $j \rightarrow \infty$ , then*

$$P \left( \ell_k(f) c_k(f) \lesssim d^{1/2} \sqrt{(\alpha - 1/2) \log k} k^{-\alpha+1/2} \right) = 1 - o(1).$$

In the case of orthogonal basis, most function will have in this class have  $c_k = k^{-\alpha+1/2}$ . Thus, Proposition 2 establishes that  $\ell_k$  is slow varying for those functions.

The following example illustrate the performance of the series estimator using different basis for a real data set.

**Example 11.** (Real Data) Here  $g(x)$  is the mean of log wage ( $y$ ) conditional on education

$$x \in \{8, 9, 10, 11, 12, 13, 14, 16, 17, 18, 19, 20\}.$$

The function  $g(x)$  is computed using population data – the 1990 Census data for the U.S. men of prime age (see Angrist, Chernozhukov and Fernandez-Val Angrist et al. (2006) for more details). We would like to know how well this function is approximated when common approximation methods are used to form the regressors. For simplicity we assume that  $w_i$  is uniformly distributed (otherwise we can weigh by the frequency). In population, least squares estimator solves the approximation problem:  $\min_b E[\{g(x_i) - p_i' b\}^2]$  for  $p_i = p(x_i)$ , where we form  $p(x)$  as (a) linear spline (Figure 2, left) and (b) Polynomial series (Figure 2, right), such that dimension of  $p(x)$  is either  $K = 3$  or  $K = 8$ .

Then we compare the function  $g(x)$  to the linear approximation  $g(x)' \beta$  graphically. We also record RMSAE as well as the maximum error MAE. The approximation errors are given in the following table:

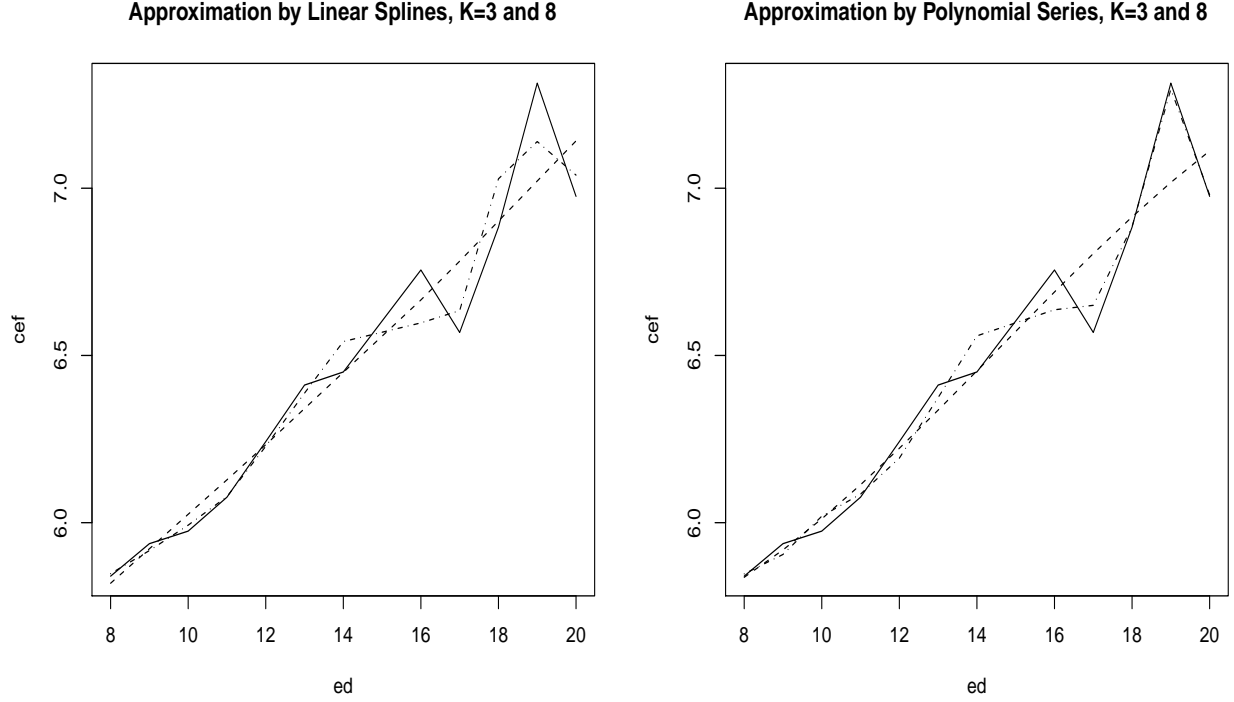


FIGURE 1.

	spline $K = 3$	spline $K = 8$	Poly $K = 3$	Poly $K = 8$
$L_2$ Error	0.12	0.08	0.12	0.05
$L_\infty$ Error	0.29	0.17	0.30	0.12

In this example, the Lebesgue constant of the polynomial approximations is comparable to the Lebesgue constant of the spline approximations.

#### 4. LIMIT THEORY

**4.1.  $L^2$  rate of convergence.** After we have established the set-up, we proceed to derive our results. We start with  $L^2$  rate of convergence result.

**Theorem 1** ( $L^2$  rate of convergence). *Assume that conditions A.1-A.3 hold. Then, under  $c_k \searrow 0$  we have*

$$\|g - \hat{g}\|_{F,2} \lesssim_P \sqrt{k/n} + c_k$$

and under  $c_k \searrow c_\infty > 0$  we have

$$\|g - \hat{g}\|_{F,2} - c_\infty \lesssim_P \sqrt{k/n} + c_k - c_\infty + \sqrt{\frac{k \log n}{n}} \cdot \ell_k c_k.$$

For most series with  $\xi_k \lesssim \sqrt{k}$  the condition  $\xi_k^2 \log n = o(n)$  amounts to  $k \log n = o(n)$ . This result weakens the rate requirements obtained in (Newey, 1997; Huang, 2003a) with unknown design and is as sharp as the result of Stone (1994) obtained for splines only. Under correct specification, the fastest rate is achieved by setting the approximation error and the sampling error to be of the same order,

$$\sqrt{k/n} \asymp c_k.$$

One consequence of this results is for the common  $\alpha$ -smooth classes the series estimators achieve the optimal rate of convergence in the  $L^2$  metric under very weak assumptions.

**4.2. Pointwise Limit Theory.** Next we focus on pointwise limit theorems. That is, for a fix sequence  $\{\alpha_i\}$ ,  $\sum_i \alpha_i^2 = 1$ . As we will show, pointwise results can be achieved under weak conditions similarly to the ones we required to achieve the rates of convergence in Theorem 1.

**Lemma 1** (Pointwise Linearization). *Suppose A.1-A.3 hold. We have that for any  $\alpha \in S^{k-1}$*

$$\sqrt{n}\alpha'(\hat{\beta} - \beta) = \alpha' \mathbb{G}_n[p_i(\epsilon_i + r_i)] + R_{1n}, \quad (4.5)$$

where the term  $R_{1n}$ , summarizing the impact of unknown design, obeys

$$R_{1n} \lesssim_P \sqrt{\frac{\xi_k^2 \log n}{n}} (1 + \sqrt{k} \ell_k c_k). \quad (4.6)$$

Moreover,

$$\sqrt{n}\alpha'(\hat{\beta} - \beta) = \alpha' \mathbb{G}_n[p_i \epsilon_i] + R_{1n} + R_{2n}, \quad (4.7)$$

where the term  $R_{2n}$ , summarizing the impact of approximation error on the sampling error of the estimator, obeys

$$R_{2n} \lesssim_P \ell_k c_k. \quad (4.8)$$

We obtain this linearization and subsequent pointwise normality results under considerably weaker conditions on the growth of  $k$  than those published in the literature, which typically impose that

$$k \xi_k^2 / n \rightarrow 0,$$

whereas here

$$\xi_k^2 \log n/n \rightarrow 0$$

is made possible in many cases. Also, as a special case, we recover the extremely sharp results of Stone and Huang for splines, who do not impose  $k\xi_k^2/n \rightarrow 0$  under the condition that maximal approximation error  $c_k \ell_k$  vanishes at  $\sqrt{k \log n}$  rate, albeit here we generally do not require  $\sqrt{k \log n} \ell_k c_k \rightarrow 0$ , so our results for this special case are slightly more general. However, as in Stone and Huang, our conditions on the growth of  $k$  are the weakest when that maximal approximation error  $c_k \ell_k$  vanishes at  $\sqrt{k \log n}$  rate. In summary, the only condition that generally matters for linearization (4.5) is that  $R_{1n} \rightarrow 0$ . In particular, our results in (4.5)-(4.6) allow for misspecification, albeit in this case, the requirement  $R_{1n} \rightarrow 0$  limits the growth of  $k$ . Moreover, we conjecture that the bound on  $R_{1n}$  can be improved for splines to

$$R_{1n} \lesssim_P \sqrt{\frac{\xi_k^2 \log n}{n}} (1 + \sqrt{\log n} \cdot \ell_k c_k).$$

since it is attained by local polynomials and splines are also similarly localized.

In order to establish normality we require the following technical conditions.

**Assumption A.4.** *The disturbance  $\epsilon_i$  is conditionally uniformly integrable, namely for each  $M \rightarrow \infty$ ,*

$$\sup_{x \in \mathcal{X}} E[\epsilon_i^2 1\{|\epsilon_i| > M\} | x_i = x] \rightarrow 0$$

*and the maximal approximation error is not too large*

$$\sup_{x \in \mathcal{X}} |r(x)| \leq \ell_k c_k = o(\sqrt{n/\xi_k}).$$

**Theorem 2** (Pointwise Normality). *Consider our least squares problem or, more generally, any problem where the estimator of  $g(x) = p(x)' \beta + r(x)$  takes the form  $p(x)' \hat{\beta}$ , where  $\hat{\beta}$  admits linearization of the form (4.5)-(4.8). Suppose A.1-A.4 hold.*

*We have that if  $R_{1n} \rightarrow_P 0$ , for any deterministic sequence  $\{\alpha\}$  with  $\|\alpha\| = 1$*

$$\sqrt{n} \frac{\alpha'(\hat{\beta} - \beta)}{\|\alpha' \Omega^{1/2}\|} =_d N(0, 1) + o_P(1),$$

*where under  $R_{2n} \not\rightarrow_P 0$ , we set  $\Omega = \tilde{\Omega} := Q^{-1} E[(\epsilon_i + r_i)^2 p_i p_i'] Q^{-1}$ , and under  $R_{2n} \rightarrow_P 0$  we can set  $\Omega = \Omega_0 := Q^{-1} E[\epsilon_i^2 p_i p_i'] Q^{-1}$ . Moreover, for any deterministic sequence  $x \in \mathcal{X}$  and*

$$s(x) := \Omega^{1/2} p(x)$$

$$\sqrt{n} \frac{p(x)'(\hat{\beta} - \beta)}{\|s(x)\|} =_d N(0, 1) + o_P(1),$$

and if the approximation error is negligible relative to the standard error, namely  $\sqrt{nr}(x) = o(\|s(x)\|)$ ,

$$\sqrt{n} \frac{(\hat{g}(x) - g(x))}{\|s(x)\|} =_d N(0, 1) + o_P(1).$$

The result delivers pointwise convergence in distribution uniformly in  $x \in \mathcal{X}$  since  $\mathcal{X}$  is compact and we allowed for any deterministic sequence within  $\mathcal{X}$ . The comments given after the linearization result in Lemma 1 apply here as well. Note that the normalization factor  $\|s(x)\|$  is the pointwise standard error, and it is of a typical order  $\|s(x)\| \propto \sqrt{k}$  at most points. (For splines and trigonometric series this holds uniformly across all points.) In this case the condition for negligibility of approximation error  $\sqrt{nr}(x)/\|s(x)\| \rightarrow 0$  can be replaced by

$$\sqrt{n/k} \cdot c_k \ell_k \rightarrow 0.$$

**4.3. Uniform Limit Theory.** Finally we turn to a uniform limit theory. Not surprising, stronger conditions are required for our results to hold when compared to the pointwise case. Here we need the following assumption on the tails of the regressor errors and on the basis.

**Assumption A.5** For some  $m > 2$ ,

$$\sup_{x \in \mathcal{X}} E[|\epsilon_i|^m | x_i = x] < \infty \quad \text{and} \quad \frac{\xi_k^{2m/(m-2)} \log n}{n} \lesssim 1.$$

Letting  $\alpha(x) := p(x)/\|p(x)\|$ , there is a constant  $a < \infty$  such that for all  $x, x' \in \mathcal{X}$

$$\|\alpha(x) - \alpha(x')\| \leq L_{1k} \|x - x'\|, \quad L_{1k} \lesssim k^a.$$

**Lemma 2** (Uniform Linearization). *Suppose that Assumptions A.1-A.5 are satisfied. Then, uniformly in  $x \in \mathcal{X}$*

$$\sqrt{n} \alpha(x)'(\hat{\beta} - \beta) = \alpha(x)' \mathbb{G}_n[p_i(\epsilon_i + r_i)] + R_{1n}, \quad (4.9)$$

where  $R_{1n}$ , summarizing the impact of unknown design, obeys

$$R_{1n} \lesssim_P \sqrt{\frac{\xi_k^2 \log n}{n}} (n^{1/m} \sqrt{\log n} + \sqrt{k} \cdot \ell_k c_k). \quad (4.10)$$

Moreover,

$$\sqrt{n}\alpha(x)'(\hat{\beta} - \beta) = \alpha(x)'\mathbb{G}_n[p_i\epsilon_i] + R_{1n} + R_{2n}, \quad (4.11)$$

where  $R_{2n}$ , summarizing the impact of approximation error on the sampling error of the estimator, obeys

$$R_{2n} \lesssim_P \sqrt{\log n} \cdot \ell_k c_k. \quad (4.12)$$

We obtain this linearization under weak conditions (in fact, it is not clear if anyone has gotten any analogous results before) also allowing for non-vanishing approximation error.

**Theorem 3** (Uniform Rate). *Consider our least squares problem or, more generally, any problem where the estimator of  $g(x) = p(x)'\beta + r(x)$  takes the form  $p(x)'\hat{\beta}$ , where  $\hat{\beta}$  admits uniform linearization of the form (4.9)-(4.12).*

*Under Assumptions A.1-A.5, we have that*

$$\sup_{x \in \mathcal{X}} |\alpha(x)'\mathbb{G}_n[p_i\epsilon_i]| \lesssim_P \sqrt{\log n}.$$

Moreover, for  $R_{1n}$  and  $R_{2n}$  given above we have

$$\sup_{x \in \mathcal{X}} |p(x)'(\hat{\beta} - \beta)| \lesssim_P \frac{\xi_k}{\sqrt{n}} (\sqrt{\log n} + R_{1n} + R_{2n})$$

and

$$\sup_{x \in \mathcal{X}} |\hat{g}(x) - g(x)| \lesssim_P \frac{\xi_k}{\sqrt{n}} (\sqrt{\log n} + R_{1n} + R_{2n}) + \ell_k c_k.$$

The resulting rates are close to the optimal rate within logs if the Lebesgue constant  $\ell_k$  behaves like  $\log n$ , which is reasonable in a number of examples, and if  $R_{1n} + R_{2n} \lesssim_P \log^c n$ , which is possible in many though not all cases. Again, conditions here improve the rates obtained in the previous work of (Newey, 1997). Relative to pointwise or  $L^2$  results, we get only an extra  $\log n$  factor in the rate. Note, however, that the assumptions on the error term are much stronger than in the pointwise case. If the errors have heavier tails, then the uniform rates can be much slower. In such cases, if one is simply interested in estimates of some location function, then one could use median regression estimator that will achieve faster uniform convergence rates, since the “errors” in the linearized version of this estimator are just Bernoulli and therefore are sub-Gaussian.

The following result is an extension of the result obtained by Chernozhukov et al. (2009); unlike their result, this result allows for a non-vanishing specification error. In particular, we make a distinction between  $\tilde{\Omega} := Q^{-1}E[(\epsilon_i + r_i)^2 p_i p_i']Q^{-1}$ , and  $\Omega_0 := Q^{-1}E[\epsilon_i^2 p_i p_i']Q^{-1}$  which are potentially asymptotically different if  $R_{2n} \not\rightarrow_P 0$ .

**Theorem 4** (Strong Approximation by a Gaussian Process). *Consider our least squares problem or, more generally, any problem where the estimator of  $g(x) = p(x)'\beta + r(x)$  takes the form  $p(x)'\hat{\beta}$ , where  $\hat{\beta}$  admits uniform linearization of the form (4.9)-(4.12). Suppose that A.1-A.3 hold, A.5 hold with  $m \geq 3$ , and that  $R_{1n} = o_P(a_n^{-1})$ , where for purposes of an application later we need  $a_n = \log n$ , and that*

$$a_n^6 k^4 \xi_k^2 (1 + \ell_k^3 c_k^3)^2 \log^2 n/n \rightarrow 0.$$

*Then for some  $\mathcal{N}_k \sim N(0, I_k)$ ,*

$$\sqrt{n} \frac{\alpha'(\hat{\beta} - \beta)}{\|\alpha'\Omega^{1/2}\|} =_d \frac{\alpha'\Omega^{1/2}}{\|\alpha'\Omega^{1/2}\|} \mathcal{N}_k + o_P(a_n^{-1}) \text{ in } \ell^\infty(S^{k-1})$$

*as stochastic processes indexed by  $\alpha \in S^{k-1}$ , so that for  $s(x) = [p(x)'\Omega^{1/2}]'$*

$$\sqrt{n} \frac{p(x)'(\hat{\beta} - \beta)}{\|s(x)\|} =_d \frac{s(x)'}{\|s(x)\|} \mathcal{N}_k + o_P(a_n^{-1}) \text{ in } \ell^\infty(\mathcal{X}),$$

*and if  $\sup_{x \in \mathcal{X}} \sqrt{n}|r(x)|/\|s(x)\| = o_P(a_n^{-1})$ ,*

$$\sqrt{n} \frac{\hat{g}(x) - g(x)}{\|s(x)\|} =_d \frac{s(x)'}{\|s(x)\|} \mathcal{N}_k + o_P(a_n^{-1}) \text{ in } \ell^\infty(\mathcal{X}).$$

*Under  $R_{1n} = o_P(a_n^{-1})$ , we set  $\Omega = \tilde{\Omega}$ , and under  $R_{2n} = o_P(a_n^{-1})$  we can set  $\Omega = \Omega_0$ .*

This result is much stronger than the pointwise normality result: it asserts that the entire studentized nonparametric estimation process is uniformly close to a Gaussian process of the stated form.

Another related inference method is the weighted bootstrap. Consider a set of weights  $h_1, \dots, h_n$  that are i.i.d. draws from the standard exponential distribution. For each draw of such weights, define the weighted bootstrap draw of the least squares estimator as a solution to the least squares problem weighted by  $h_1, \dots, h_n$ , namely

$$\hat{\beta}^b \in \arg \min_{b \in \mathbb{R}^k} \mathbb{E}_n[h_i(y_i - p_i'b)^2].$$

The following theorem establishes that the weighted bootstrap distribution is valid for approximating the distribution of the least squares estimator.

**Theorem 5** (Weighted Bootstrap Method). *(1) Suppose that A.1-A.5 hold and  $(n^{2/m} \log n + k \ell_k^2 c_k^2) \xi_k^2 \log^6 n = o(n)$ . Then the weighted bootstrap process satisfies*

$$\sqrt{n} \alpha(x)'(\hat{\beta}^b - \hat{\beta}) = \alpha(x)' \mathbb{G}_n[(h_i - 1)p_i(\epsilon_i + r_i)] + R_{1n},$$



where

$$\sup_{x \in \mathcal{X}} |R_{1n}| \lesssim_P \sqrt{\frac{\xi_k^2 \log^4 n}{n}} (n^{1/m} \sqrt{\log n} + \sqrt{k} \ell_k c_k) = o(1/\log n).$$

The bound continues to hold in  $P$ -probability if we replace the unconditional probability  $P$  by the conditional probability  $P^*(\cdot|X)$ .

(2) Furthermore, under the conditions of Theorem 4, the weighted bootstrap process indexed by  $\alpha \in S^{k-1}$  approximates some Gaussian process  $\mathcal{N}_k \sim N(0, I_k)$  defined in Theorem 4, that is:

$$\|\Omega^{-1/2} \sqrt{n}(\hat{\beta}^b - \hat{\beta}) - \mathcal{N}_k\| = o_P(1/\log n).$$

We close this section by establishing sufficient conditions for the consistent estimation of  $\Omega$ .

**Theorem 6** (Matrices Estimation). *Let  $Q = E[p_i p_i']$  and  $\Sigma = E[(\epsilon_i + r_i)^2 p_i p_i']$ . Assume that  $v_n^2 = E[\max_{1 \leq i \leq n} |\epsilon_i|^2]$  is such that  $(1 + \ell_k^2 c_k^2 k \log n + v_n^2) \xi_k^2 \log^2 n = o_P(n)$ . Under A.1-A.2, for  $\hat{Q} = E_n[p_i p_i']$  and  $\hat{\Sigma} = E_n[\hat{\epsilon}_i^2 p_i p_i']$ , where  $\hat{\epsilon}_i = y_i - p_i' \hat{\beta}$ , we have we have*

$$\|Q - \hat{Q}\| \lesssim_P \sqrt{\frac{\xi_k^2 \log n}{n}} \|Q\| = o(1) \quad \text{and} \quad \|\Sigma - \hat{\Sigma}\| \lesssim_P (\|Q\| \vee \|\Sigma\|)(v_n + \ell_k c_k) \sqrt{\frac{\xi_k^2 \log^2 n}{n}}.$$

Moreover, under these conditions if for some sequence  $a_n \rightarrow \infty$ ,  $\|Q - \hat{Q}\| = o_P(1/a_n)$  and  $\|\Sigma - \hat{\Sigma}\| = o_P(1/a_n)$ , we have  $\|\Omega - \hat{\Omega}\| = o_P(1/a_n)$  where  $\hat{\Omega} = \hat{Q}^{-1} \hat{\Sigma} \hat{Q}^{-1}$ .

In the case of a bounded basis, Theorem 6 allows for consistent estimation of the matrix  $Q$  under the mild condition  $k \log n = o(n)$ . Not surprising, the estimation of  $\Sigma$  depends on the tail behavior of the error term. We note that under condition A.5, we have  $v_n^2 \lesssim n^{2/m}$ .

## 5. RATES AND INFERENCE ON LINEAR FUNCTIONALS

In this section derive rates and inference results for linear functionals  $\theta(w), w \in \mathcal{I}$  of the conditional expectation function such as its derivative or average derivative. By the linearity of the series approximations, the linear functionals can be seen as linear functions of the least squares coefficients  $\beta$  up to an approximation error, that is

$$\theta(w) = \ell(w)' \beta + r_n(w), \quad w \in \mathcal{I},$$

where  $\ell(w)' \beta$  is the series approximation, with  $\ell(w)$  denoting the  $k$ -vector of loadings on the coefficients, and  $r_n(w)$  is the remainder term, which corresponds to the approximation error.

In order to perform inference, we construct estimators for  $\sigma_n^2(w) = \ell(w)' \Omega \ell(w)/n$ , the variance of the associated linear functionals, as

$$\hat{\sigma}_n^2(w) = \ell(w)' \hat{\Omega} \ell(w)/n. \quad (5.13)$$

By Theorem 6, under our conditions (5.13) is uniformly consistent for  $\sigma_n^2(w)$ , namely  $\hat{\sigma}_n^2(w)/\sigma_n^2(w) = 1 + o_P(1)$  uniformly over  $w \in \mathcal{I}$ .

**5.1. Pointwise Results for Linear Functionals.** Next we state regularity conditions on the loadings and approximation errors associated with the linear functional  $\theta$ .

**Condition P.**

P.1 *The approximation error is small, namely  $\sqrt{n}|r_n(w)|/\|\ell(w)\| = o(1)$ .*

P.2 *The norm of the loading  $\ell(w)$  satisfies:  $\|\ell(w)\| \lesssim \xi_\theta(k, w)$ .*

**Theorem 7** (Pointwise Convergence Rate for Linear Functionals). *Assume that the conditions of Theorem 2 and Condition P hold, then*

$$|\hat{\theta}(w) - \theta(w)| \lesssim_P \frac{\xi_\theta(k, w)}{\sqrt{n}}. \quad (5.14)$$

To perform inference, we consider the t-statistic:

$$t_n(w) = \frac{\hat{\theta}(w) - \theta(w)}{\hat{\sigma}_n(w)}.$$

Under Condition P, the approximation error is small, so that

$$t_n(w) = \frac{\ell(w)'(\hat{\beta} - \beta)}{\hat{\sigma}_n(w)} + o_P(1).$$

We can carry out standard inference based on this statistic because  $t_n(w) \rightarrow_d N(0, 1)$ .

**Theorem 8** (Pointwise Inference for Linear Functionals). *Suppose that the conditions of Theorem 2, Theorem 6, and Condition P hold. Then,*

$$t_n(w) \rightarrow_d N(0, 1).$$

**5.2. Uniform Results for Linear Functionals.** In studying uniform rates and inference we use the sup norm over the indices of interest  $\mathcal{I} \subset \mathbb{R}^d$ , namely, for  $f : \mathcal{I} \mapsto \mathbb{R}^m$ , define the norm

$$\|f\|_{\mathcal{I}} := \sup_{w \in \mathcal{I}} |f(w)|.$$

We shall invoke the following assumptions to establish rates and uniform inference results over the region  $\mathcal{I}$ .

**Condition U.**

- U.1 *The approximation error is small, namely  $\sqrt{n} \log n \sup_{w \in \mathcal{I}} \|r_n(w)/\|\ell(w)\|\| = o(1)$ .*
- U.2 *The loadings  $\ell(w)$  are uniformly bounded and admit Lipschitz coefficients  $\xi_\theta^L(k, \mathcal{I})$ , that is,*

$$\|\ell\|_{\mathcal{I}} \lesssim \xi_\theta(k, \mathcal{I}), \quad \|\ell(w) - \ell(w')\| \leq \xi_\theta^L(k, \mathcal{I}) \|w - w'\|, \quad \text{and}$$

$$\log[\text{diam}(\mathcal{I}) \vee \xi_\theta(k, \mathcal{I}) \vee \xi_\theta^L(k, \mathcal{I}) \vee \xi_k] \lesssim \log k.$$

The value of  $\xi_\theta(k, \mathcal{I})$  depends on the choice of basis for the series estimator and on the linear functional. Newey (1997) and Chen (2006) provides several examples. In the case of regression splines, after a possible renormalization so that  $\mathcal{X} = [-1, 1]^d$ , it has been established that  $\xi_k \lesssim \sqrt{k}$  and  $\sup_{x \in \mathcal{X}} \|\partial_x^m p(x)\| \lesssim k^{1/2+m}$  (Newey, 1997). With this basis we have for the function itself  $\xi_\theta(k, \mathcal{I}) \lesssim \sqrt{k}$  ( $\theta(x) = g(x)$  and  $\ell(x) = p(x)$ ); for the derivative  $\xi_\theta(k, \mathcal{I}) \lesssim k^{3/2}$  ( $\theta(x) = \partial_{x_j} g(x)$  and  $\ell(x) = \partial_x p(x)$ ); for the average derivative  $\xi_\theta(k) \lesssim 1$  ( $\theta = \int \partial_{x_j} g(x) d\mu(x)$ ,  $\text{supp}(\mu) \subset \text{int}\mathcal{X}$ ,  $|\partial_{x_k} \mu(x)| \lesssim 1$ ,  $\ell = \int \partial_{x_j} p(x) \mu(x) dx = - \int p(x) \partial_{x_j} \mu(x) dx$ ).

**Theorem 9** (Uniform Convergence Rate for Linear Functionals). *Assume that the conditions of Lemma 2 and Condition U hold, and  $d\xi_\theta^2(k, \mathcal{I})\xi_k^2 \log^2 n = o(n)$ , then*

$$\sup_{w \in \mathcal{I}} |\hat{\theta}(w) - \theta(w)| \lesssim_P \frac{\xi_\theta(k, \mathcal{I}) \vee 1}{\sqrt{n}} \sqrt{\log n}. \quad (5.15)$$

In this case we consider the t-statistic process:

$$\left\{ t_n(w) = \frac{\hat{\theta}(w) - \theta(w)}{\hat{\sigma}_n(w)}, \quad w \in \mathcal{I} \right\}.$$

Under our assumptions the approximation error is small, so that

$$t_n(w) = \frac{\ell(w)'(\hat{\beta} - \beta)}{\hat{\sigma}_n(w)} + o_P(1/\log n) \text{ in } \ell^\infty(\mathcal{I}).$$

The main result on inference is that the t-statistic process can be strongly approximated by the following Gaussian coupling:

$$\left\{ t_n^*(w) = \frac{\ell(w)' \hat{\Omega}^{1/2} \mathcal{N}_k / \sqrt{n}}{\hat{\sigma}_n(w)}, \quad w \in \mathcal{I} \right\}. \quad (5.16)$$

The following theorem shows that these couplings approximate the distribution of the t-statistic process in large samples.

**Theorem 10** (Strong Approximation of Inferential Processes by Gaussian Coupling). *Suppose that the conditions of Theorems 4, Theorem 6, and Condition U hold. Then,*

$$t_n(w) =_d t_n^*(w) + o_P(1/\log n), \text{ in } \ell^\infty(\mathcal{I}).$$

To construct uniform two-sided confidence bands for  $\{\theta(w) : w \in \mathcal{I}\}$ , we consider the maximal  $t$ -statistic

$$\|t_n\|_{\mathcal{I}} = \sup_{w \in \mathcal{I}} |t_n(w)|,$$

as well as the couplings to this statistic in the form:

$$\|t_n^*\|_{\mathcal{I}} = \sup_{w \in \mathcal{I}} |t_n^*(w)|.$$

Ideally, we would like to use quantiles of the first statistic as critical values, but we do not know them. We instead use quantiles of the second statistic as large sample approximations. Let  $k_n(1 - \alpha)$  denote the  $1 - \alpha$  quantile of random variable  $\|t_n^*\|_{\mathcal{I}}$  conditional on the data  $\mathcal{D}_n$ , i.e.

$$k_n(1 - \alpha) = \inf\{t : P(\|t_n^*\|_{\mathcal{I}} \leq t | \mathcal{D}_n) \geq 1 - \alpha\}.$$

This quantity can be computed numerically by Monte Carlo methods, as we illustrate in the empirical section.

Let  $\delta_n > 0$  be a finite sample expansion factor such that  $\delta_n \log^{1/2} n \rightarrow 0$  but  $\delta_n \log n \rightarrow \infty$ . For example, we recommend to set  $\delta_n = 1/(4 \log^{3/4} n)$ . Then for  $c_n(1 - \alpha) = k_n(1 - \alpha) + \delta_n$  we define the confidence bands of asymptotic level  $1 - \alpha$  to be

$$[i(w), \bar{i}(w)] = [\hat{\theta}(w) - c_n(1 - \alpha)\hat{\sigma}_n(w), \hat{\theta}(w) + c_n(1 - \alpha)\hat{\sigma}_n(w)], \quad w \in \mathcal{I}.$$

The following theorem establishes the asymptotic validity of these confidence bands. The last result relies on the additional property of anti-concentration. The anti-concentration property holds if, after appropriate scaling by some deterministic sequences  $a_n$  and  $b_n$ , the inferential statistic  $a_n(\|t_n\|_{\mathcal{I}} - b_n)$  has a continuous limit distribution. More generally, it holds if for any subsequence of integers  $\{n_k\}$  there is a further subsequence  $\{n_{k_r}\}$  along which  $a_{n_{k_r}}(\|t_{n_{k_r}}\|_{\mathcal{I}} - b_{n_{k_r}})$  has a continuous limit distribution, possibly dependent on the subsequence. We expect anti-concentration to hold in our case, but our constructions and results do not critically hinge on it.

**Theorem 11** (Uniform Inference for Linear Functionals). *Suppose that the conditions of Theorem 10 hold.*

(1) *Then*

$$P\left\{\|t_n\|_{\mathcal{I}} \leq c_n(1 - \alpha)\right\} \geq 1 - \alpha + o(1). \quad (5.17)$$

(2) As a consequence, the confidence bands constructed above cover  $\theta(w)$  uniformly for all  $w \in \mathcal{I}$  with probability that is asymptotically no less than  $1 - \alpha$ , namely

$$P\left\{\theta(w) \in [i(w), \ddot{i}(w)], \text{ for all } w \in \mathcal{I}\right\} \geq 1 - \alpha + o(1). \quad (5.18)$$

(3) The width of the confidence band  $2c_n(1 - \alpha)\hat{\sigma}_n(w)$  obeys uniformly in  $w \in I$ :

$$2c_n(1 - \alpha)\hat{\sigma}_n(w) = 2k_n(1 - \alpha)(1 + o_P(1))\sigma_n(w). \quad (5.19)$$

(4) Furthermore, if  $\|t_n^*\|_{\mathcal{I}}$  does not concentrate at  $k_n(1 - \alpha)$  at a rate faster than  $\sqrt{\log n}$ , that is, it obeys the anti-concentration property  $P(\|t_n^*\|_{\mathcal{I}} \leq k_n(1 - \alpha) + \varepsilon_n) = 1 - \alpha + o(1)$  for any  $\varepsilon_n = o(1/\sqrt{\log n})$ , then the inequalities in (5.17) and (5.18) hold as equalities, and the finite sample adjustment factor  $\delta_n$  could be set to zero.

Theorem 11 shows that the confidence bands constructed above maintain the required level asymptotically and establishes that the uniform width of the bands is of the same order as the uniform rate of convergence. Moreover, confidence intervals are asymptotically similar under anti-concentration .

A similar strategy was proposed in Chernozhukov et al. (2009) for inference on the minimum of a function. Since the limit distribution may not exist, the insight was to use distributions provided by couplings. Because the limit distribution does not necessarily exist, it is not immediately clear that the confidence intervals maintain the right asymptotic level. However, the additional adjustment factor  $\delta_n$  assures the right asymptotic level. A potential downside for using the adjustment  $\delta_n$  is that the confidence intervals may not be similar, i.e. remain asymptotically conservative in coverage. However, the width of the confidence intervals is not asymptotically conservative, since  $\delta_n$  is negligible compared to  $k_n(1 - \alpha)$ . Nonetheless, if the anti-concentration property holds, then the confidence intervals automatically become asymptotically similar.

## 6. TOOLS: MAXIMAL INEQUALITIES FOR MATRICES AND EMPIRICAL PROCESSES

In this section we collect the main technical tools that our analysis rely upon, namely Khinchin Inequalities for Matrices and a Data Dependent Maximal Inequalities.

**6.1. Khinchin Inequalities for Matrices.** Consider the Schatten norm  $S_P$  on symmetric  $k \times k$  matrices  $Q$  as

$$\|Q\|_{S_P} = \left( \sum_{j=1}^k |\lambda_j(Q)|^p \right)^{1/p}.$$

The case  $p = \infty$  recovers the operator norm  $\|\cdot\|$  and  $p = 2$  the Frobenius norm. It is obvious that for any  $p \geq 1$

$$\|Q\| \leq \|Q\|_{S_P} \leq k^{1/p} \|Q\|.$$

Therefore, setting  $p = \log k$ , we get equivalence

$$\|Q\| \leq \|Q\|_{S_{\log k}} \leq e \|Q\|. \quad (6.20)$$

**Lemma 3** (Khinchin Inequality for Matrices). *For symmetric  $k \times k$ -matrices  $Q_i$ ,  $i = 1, \dots, n$ , and  $2 \leq p < \infty$ , and an i.i.d. sequence of Rademacher variables  $\varepsilon_1, \dots, \varepsilon_n$  we have*

$$a_P \left\| (\mathbb{E}_n[Q_i^2])^{1/2} \right\|_{S_P} \leq \left( E_\varepsilon \|\mathbb{G}_n[\varepsilon_i Q_i]\|_{S_P}^p \right)^{1/p} \leq b_P \left\| (\mathbb{E}_n[Q_i^2])^{1/2} \right\|_{S_P}$$

where

$$b_P \leq [2^{1/2} \pi / e]^{1/2} \cdot \sqrt{p}.$$

As a consequence of equivalence (6.20) if  $k \geq e^2$  we have

$$E_\varepsilon \|\mathbb{G}_n[\varepsilon_i Q_i]\| \lesssim \sqrt{\log k} \|(\mathbb{E}_n[Q_i^2])^{1/2}\|$$

The notable feature of this inequality is the  $\sqrt{\log k}$  factor instead of the  $\sqrt{k}$  factor expected from the conventional maximal inequalities based on entropy. This inequality due to Lust-Picard and Pisier (1991) generalizes the Khinchin inequality for vectors. A version of this inequality was derived by Guédon and Rudelson (2007) using generalized entropy (majorizing measure) arguments. This is another striking example where the use of generalized entropy yields drastic improvements over the use of entropy. Prior to this Talagrand (1996a) provided ellipsoidal examples where the difference between the two approaches was even more extreme.

**6.2. LLN for Matrices.** The following lemma is a variant of a fundamental result obtained by Rudelson (1999).

**Lemma 4** (Matrix LLN). *Let  $Q_1, \dots, Q_n$  be i.n.i.d. symmetric non-negative  $k \times k$ -matrices with  $k \geq e^2$  such that  $Q = \mathbb{E}_n[E[Q_i]]$  and  $\|Q_i\| \leq M$  a.s., then for  $\widehat{Q} = \mathbb{E}_n[Q_i]$*

$$\Delta := E \|\widehat{Q} - Q\| \lesssim \sqrt{\frac{M(1 + \|Q\|) \log n}{n}}.$$

*In particular, if  $Q_i = p_i p_i'$ , with  $\|p_i\| \leq \xi_k$  a.s., then*

$$\Delta := E \|\widehat{Q} - Q\| \lesssim \sqrt{\frac{\xi_k^2(1 + \|Q\|) \log n}{n}}.$$

**6.3. Maximal Inequalities.** Consider a function class  $\mathcal{F}$  collecting functions mapping some set  $\mathcal{Z}$  to  $\mathbb{R}$ , equipped with an envelope function  $F(z) \geq \sup_{f \in \mathcal{F}} |f(z)|$ . The *covering number*  $N(\varepsilon, \mathcal{F}, L^2(Q))$  is the minimal number of  $L^2(Q)$ -balls of radius  $\varepsilon$  needed to cover the function set  $\mathcal{F}$ . The *covering number* relative to the envelope function is given by

$$N\left(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L^2(Q)\right). \quad (6.21)$$

The *entropy* is the logarithm of the covering number.

We rely on the following result.

**Proposition 3.** *Let  $(\epsilon_1, X_1), \dots, (\epsilon_n, X_n)$  be i.i.d. random vectors in  $\mathbb{R}^{d+1}$  with  $E[\epsilon_i | X_i] = 0$  and  $\sigma^2 := \sup_x E[\epsilon_i^2 | X_i = x] < \infty$ . Let  $\mathcal{F}$  be a class of functions on  $\mathbb{R}^d$  such that  $E[f(X_1)^2] = 1$  (normalization) and  $\|f\|_\infty \leq b$  for all  $f \in \mathcal{F}$ . Let  $\mathcal{G} := \{(\epsilon, x) \in \mathbb{R}^{d+1} \mapsto \epsilon f(x) : f \in \mathcal{F}\}$ . Suppose that there exist constants  $A > e^2$  and  $V \geq 2$  such that*

$$\sup_Q N(\mathcal{G}, L^2(Q), \epsilon \|G\|_{L^2(Q)}) \leq (A/\epsilon)^V$$

for all  $0 < \epsilon \leq 1$  for the envelope  $G(\epsilon, x) := |\epsilon|b$ . If for some  $m > 2$   $E[|\epsilon_1|^m] < \infty$ , then

$$E \left[ \left\| \sum_{i=1}^n \epsilon_i f(X_i) \right\|_{\mathcal{F}} \right] \leq C \left[ (\sigma + \sqrt{E[|\epsilon_1|^m]}) \sqrt{nV \log(Ab)} + Vb^{m/(m-2)} \log(Ab) \right],$$

where  $C$  is a universal constant.

The proof is based on a truncation argument and maximal inequalities for uniformly bounded classes of functions developed in Giné and Koltchinskii (2006). We recall its version.

**Theorem 12** (Giné and Koltchinskii (2006)). *Let  $\xi_1, \dots, \xi_n$  be i.i.d. random variables taking values in a measurable space  $(S, \mathcal{S})$  with common distribution  $P$ . Let  $\mathcal{F}$  be a suitably measurable class of functions on  $S$  with envelope  $F$ . Let  $\sigma^2$  be a constant such that  $\sup_{f \in \mathcal{F}} \text{var}(f) \leq \sigma^2 \leq \|F\|_{L^2(P)}^2$ . Suppose that there exist constants  $A > e^2$  and  $V \geq 2$  such that  $\sup_Q N(\mathcal{F}, L^2(Q), \epsilon \|F\|_{L^2(Q)}) \leq (A/\epsilon)^V$  for all  $0 < \epsilon \leq 1$ . Then,*

$$E \left[ \left\| \sum_{i=1}^n \{f(\xi_i) - E[f(\xi_1)]\} \right\|_{\mathcal{F}} \right] \leq C \left[ \sqrt{n\sigma^2 V \log \frac{A\|F\|_{L^2(P)}}{\sigma}} + V\|F\|_\infty \log \frac{A\|F\|_{L^2(P)}}{\sigma} \right],$$

where  $C$  is a universal constant.

## APPENDIX A. PROOFS

## A.1. Proofs of Sections 2 and 3.

*Proof of Proposition 1.* For any  $\gamma$ ,  $\int (\gamma' p)^2 dF = \int (\gamma' p)^2 (dF/d\mu) d\mu$ . So that if  $dF/d\mu$  is bounded above and away from zero, the result follows since the basis is orthonormal under  $(\mathcal{X}, \mu)$ .  $\square$

*Proof of Proposition 2.* For a function  $f = \sum_{j \geq 1} p_j(x) j^{-\alpha} \xi_j \in \mathcal{F}(\alpha)$  let  $A_f(x) := \sum_{j \geq k+1} p_j(x) j^{-\alpha} \xi_j$  and

$$\bar{v}_k := d^{1/2} \sqrt{(\alpha - 1/2) \log k} k^{-\alpha+1/2}.$$

Then, the statement of the lemma is equivalent to

$$P \left( \sup_{x \in \mathcal{X}} |A_f(x)| \lesssim \bar{v}_k \right) = 1 - o(1).$$

Consider an  $\epsilon$ -net  $\mathcal{N}_\epsilon$  for  $\mathcal{X}$ , and for some  $L \geq 1$  let  $\mathcal{H}_k := \{f \in \mathcal{F} : |A_f(x) - A_f(x')| \leq L \|x - x'\| \text{ for all } x, x' \in \mathcal{X}\}$ . Then

$$\begin{aligned} P(\sup_{x \in \mathcal{X}} |A_f(x)| \gtrsim \bar{v}_k) &\leq P(f \notin \mathcal{H}_k) + P(f \in \mathcal{H}_k, \sup_{x \in \mathcal{N}_\epsilon} |A_f(x)| \gtrsim \bar{v}_k - L\epsilon) \\ &\leq P(f \notin \mathcal{H}_k) + |\mathcal{N}_\epsilon| \max_{x \in \mathcal{N}_\epsilon} P(|A_f(x)| \gtrsim \bar{v}_k - L\epsilon) \end{aligned}$$

Note that we can take  $|\mathcal{N}_\epsilon| \leq (\text{diam}(\mathcal{X})/\epsilon)^d$  and

$$\begin{aligned} E[A_f(x)^2 | x] &= E[(\sum_{j \geq k+1} j^{-\alpha} p_j(x) \xi_j)^2 | x] = E[\sum_{j \geq k+1} j^{-2\alpha} p_j^2(x) \xi_j^2 | x] \\ &\leq k^{-2\alpha+1} \sup_{j \geq k+1} p_j^2(x). \end{aligned}$$

Thus, setting  $\epsilon = k^{-\alpha+1/2}/L$  and  $\bar{v}_k := \sqrt{d \log(\text{diam}(\mathcal{X})/\epsilon)} k^{-\alpha+1/2}$  we have  $L\epsilon \lesssim \bar{v}_k$  and since  $A_f(x) \sim N(0, E[A_f(x)^2 | x])$  we have

$$P \left( f \in \mathcal{H}_k, \sup_{x \in \mathcal{N}_\epsilon} |A_f(x)| \gtrsim \bar{v}_k - L\epsilon \right) = o(1).$$

Next, to bound  $P(f \notin \mathcal{H}_k)$ , note that  $f$  is  $L$ -Lipschitz if

$$Z := \sup_{x, x' \in \mathcal{X}} \left| \frac{\sum_{j \geq k+1} \{p_j(x) - p_j(x')\} j^{-\alpha} \xi_j}{\|x - x'\|} \right| \leq L.$$

Since  $\sup_{x \in \mathcal{X}, j \geq k+1} \|\nabla p_j(x)\| \leq M_j$ , we have that for  $\delta \in (0, 1)$

$$\begin{aligned} P(Z > \sum_{j \geq k+1} j^{-\alpha} M_j \sqrt{2 \log(2j^2/\delta)}) &\leq P(\exists j \geq k+1 : |\xi_j| \geq \sqrt{2 \log(2j^2/\delta)}) \\ &\leq \sum_{j \geq k+1} \delta/j^2 \leq \delta. \end{aligned}$$

Thus, the result follows by noting that

$$\log(\text{diam}(\mathcal{X})/\epsilon) = \log(L \text{diam}(\mathcal{X}) k^{\alpha-1/2}) \lesssim \log k$$



provided we choose  $\delta = o(1)$  so that  $j^{-\alpha} M_j \sqrt{2 \log(2j^2/\delta)} = o(1)$  which leads to

$$\sum_{j \geq k+1} j^{-\alpha} M_j \sqrt{2 \log(2j^2/\delta)} \lesssim 1.$$

□

## A.2. Proofs of Section 4.1.

*Proof of Theorem 1.* We have that

$$\|g - \hat{g}\|_{F,2} \leq \|g - p'\beta\|_{F,2} + \|p'\beta - p'\hat{\beta}\|_{F,2} \leq c_k + \|p'\beta - p'\hat{\beta}\|_{F,2}$$

where under the normalization  $Q = E[p(x)p(x)'] = I$  we have

$$\|p'\beta - p'\hat{\beta}\|_{F,2} = \left[ \int (\beta - \hat{\beta})' p(x) p(x)' (\beta - \hat{\beta}) dF(x) \right]^{1/2} = \|\hat{\beta} - \beta\|.$$

To prove the result we need to show  $\|\hat{\beta} - \beta\| \lesssim_P \sqrt{k/n}$ . We have

$$\|\hat{\beta} - \beta\| = \|\hat{Q}^{-1} \mathbb{E}_n[p_i(\epsilon_i + r_i)]\| \leq \|\hat{Q}^{-1} \mathbb{E}_n[p_i \epsilon_i]\| + \|\hat{Q}^{-1} \mathbb{E}_n[p_i r_i]\|.$$

By the Matrix LLN of Lemma 4, which is the critical step, we have that

$$\|\hat{Q} - Q\| \rightarrow_P 0 \text{ if } \frac{\xi_k^2 \log n}{n} \rightarrow 0.$$

Therefore

$$\|\hat{Q}^{-1} \mathbb{E}_n[p_i \epsilon_i]\| \lesssim_P \|\mathbb{E}_n[p_i \epsilon_i]\| \lesssim_P \sqrt{k/n}$$

since  $\lambda_{\min}(\hat{Q}) > 1/2$  wp  $\rightarrow 1$  and by  $\sigma_i^2$  bounded

$$E[\|\mathbb{E}_n[p_i \epsilon_i]\|^2] = E[\epsilon_i^2 p_i' p_i / n] = E[\sigma_i^2 p_i' p_i / n] \lesssim E[p_i' p_i / n] = k/n.$$

Moreover when  $c_k \searrow 0$ ,

$$\|\hat{Q}^{-1} \mathbb{E}_n[p_i r_i]\| \lesssim_P \|\hat{Q}^{-1/2} \mathbb{E}_n[p_i r_i]\| \lesssim_P c_k^2$$

since  $\lambda_{\min}(\hat{Q}) > 1/2$  wp  $\rightarrow 1$ . Moreover, since  $\hat{r}_i := p_i' \hat{Q}^{-1} \mathbb{E}_n[p_i r_i]$  is a sample projection of  $r_i$  on  $p_i$ , so that

$$\|\hat{Q}^{-1/2} \mathbb{E}_n[p_i r_i]\|^2 = \mathbb{E}_n[r_i \hat{r}_i] = \mathbb{E}_n \hat{r}_i^2 \leq \mathbb{E}_n[r_i^2] \lesssim_P E[r_i^2] \leq c_k^2,$$

by the Chebyshev inequality.

Moreover when  $c_k \searrow c_\infty > 0$ ,

$$\|\hat{Q}^{-1} \mathbb{E}_n[p_i r_i]\| \leq \|\hat{Q}^{-1}\| \sup_{\|\gamma\|=1} \gamma' \mathbb{E}_n[p_i r_i] \lesssim_P \sqrt{\frac{k}{n}} \cdot \ell_k c_k,$$

where the last inequality is by Step 2 in the proof of Lemma 2.

□

### A.3. Proofs of Section 4.2.

*Proof of Lemma 1.* Decompose

$$\sqrt{n}\alpha'(\hat{\beta} - \beta) = \alpha'\mathbb{G}_n[p_i(\epsilon_i + r_i)] + \alpha'[\hat{Q}^{-1} - I]\mathbb{G}_n[p_i(\epsilon_i + r_i)].$$

We divide the proof in three steps. Step 1 and 2 establish the first linearization result. Step 3 provides the last result (a bound on  $R_{2n}$ ).

Step 1. Conditional on  $X = [x_1, \dots, x_n]$  the term

$$\alpha'[\hat{Q}^{-1} - I]\mathbb{G}_n[p_i\epsilon_i]$$

has mean zero and variance bounded by  $\alpha'[\hat{Q}^{-1} - I]\hat{Q}\bar{\sigma}^2[\hat{Q}^{-1} - I]\alpha$ . Next, since the design is random, by Matrix LLN Lemma 4 we have

$$\begin{aligned} \alpha'[\hat{Q}^{-1} - I]\hat{Q}\bar{\sigma}^2[\hat{Q}^{-1} - I]\alpha &\lesssim \bar{\sigma}^2\|\hat{Q}\|\|\hat{Q}^{-1}\|^2\|\hat{Q} - I\|^2 \\ &\lesssim_P \bar{\sigma}^2\lambda_{\max}(\hat{Q})\lambda_{\min}^{-2}(\hat{Q})\frac{\xi_k^2 \log n}{n} \\ &\lesssim_P \frac{\xi_k^2 \log n}{n} = o(1) \end{aligned}$$

since  $\xi_k^2 \log n = o(n)$ . We then conclude by Chebyshev inequality that

$$\alpha'[\hat{Q}^{-1} - I]\mathbb{G}_n[p_i\epsilon_i] \lesssim_P \sqrt{\frac{\xi_k^2 \log n}{n}}.$$

Step 2. Under the random design, we have  $E[p_i r_i] = 0$ . Thus, by Matrix LLN Lemma 4 so that  $\|\hat{Q} - I\| \lesssim_P (\xi_k^2 \log n / n)^{1/2}$  and Lemma 5 we have

$$\begin{aligned} |\alpha'(\hat{Q}^{-1} - I)\mathbb{G}_n[p_i r_i]| &\leq \|\hat{Q}^{-1} - I\| \sup_{\|\gamma\|=1} \gamma' \mathbb{G}_n[p_i r_i] \\ &\leq \|\hat{Q}^{-1}\| \|\hat{Q} - I\| \sup_{\|\gamma\|=1} \gamma' \mathbb{G}_n[p_i r_i] \\ &\lesssim_P \sqrt{\frac{\xi_k^2 \log n}{n}} \sqrt{k} \ell_k c_k, \end{aligned}$$

since

$$\begin{aligned}
E\left[\sup_{\|\gamma\|=1} |\mathbb{G}_n[\gamma' p_i r_i]|\right] &\leq \frac{1}{\sqrt{n}} E \left[ \sqrt{\sum_{j=1}^k \left( \sum_{i=1}^n p_j(x_i) r(x_i) \right)^2} \right] \\
&\leq \frac{1}{\sqrt{n}} \sqrt{\sum_{j=1}^k E \left[ \left( \sum_{i=1}^n p_j(x_i) r(x_i) \right)^2 \right]} \\
&\leq \ell_k c_k \sqrt{E[\|p(x_1)\|^2]} \wedge \xi_k c_k \leq \ell_k c_k \sqrt{k}.
\end{aligned}$$

We used that  $E[p_i r_i] = 0$  and that for any  $\|\gamma\| = 1$

$$\begin{aligned}
\left| \sum_{i=1}^n \gamma' p(x_i) r(x_i) \right| &= \left| \sum_{i=1}^n \sum_{j=1}^k \gamma_j p_j(x_i) r(x_i) \right| \\
&= \left| \sum_{j=1}^k \gamma_j \left( \sum_{i=1}^n p_j(x_i) r(x_i) \right) \right| \\
&\leq \sqrt{\sum_{j=1}^k \gamma_j^2} \sqrt{\sum_{j=1}^k \left( \sum_{i=1}^n p_j(x_i) r(x_i) \right)^2} \\
&= \sqrt{\sum_{j=1}^k \left( \sum_{i=1}^n p_j(x_i) r(x_i) \right)^2}.
\end{aligned}$$

(Note that here we cannot rely on the initial conditioning argument used in Step 1.)

Step 3. Under random design we have  $E[p_i r_i] = 0$ . Thus, the term

$$R_{2n} = \alpha' \mathbb{G}_n[p_i r_i]$$

has mean zero and variance

$$E[\alpha' p_i r_i]^2 \leq E[\alpha' p_i]^2 \ell_k^2 c_k^2 \leq \ell_k^2 c_k^2.$$

Thus, using Chebyshev inequality, this steps gives the second linearization result.  $\square$

*Proof of Theorem 2.* It suffices to prove the first result only. For any sequence  $\alpha \in S^{k-1}$ , we can write

$$\frac{\sqrt{n} \alpha'}{\|\alpha' \Omega^{1/2}\|} (\hat{\beta} - \beta) = \frac{\alpha'}{\|\alpha' \Omega^{1/2}\|} \mathbb{G}_n[p_i(\epsilon_i + r_i)] = \sum_{i=1}^n \omega_{ni}(\epsilon_i + r_i),$$

where

$$\omega_{ni} = \frac{\alpha'}{\|\alpha' \Omega^{1/2}\|} \frac{p_i}{\sqrt{n}}, \quad |\omega_{ni}| \leq \frac{\xi_k}{\sqrt{n}}, \quad |\epsilon_i + r_i| \leq |\epsilon_i| + \ell_k c_k$$

and since

$$\Omega \geq \Omega_0 \geq \underline{\sigma}^2 Q \quad (\text{A.22})$$

we also have

$$nE\|\omega_{ni}\|^2 \leq E[\alpha' p_i]^2 / \alpha' \Omega \alpha \leq 1 / \underline{\sigma}^2. \quad (\text{A.23})$$

Next we verify the Lindberg condition for the CLT. First, by construction we have

$$\text{var} \left( \sum_{i=1}^n \omega_{ni}(\epsilon_i + r_i) \right) = 1.$$

Second, for each  $\delta > 0$

$$\sum_{i=1}^n E \left[ \|\omega_{ni}\|^2 (\epsilon_i + r_i)^2 1_{\{\|\omega_{ni}\| |\epsilon_i + r_i| > \delta\}} \right] \rightarrow 0,$$

since the left hand side is bounded by

$$2nE \left[ \|\omega_{ni}\|^2 \epsilon_i^2 1_{\{|\epsilon_i| + \ell_k c_k > \delta / \|\omega_{ni}\|\}} \right] + nE \left[ \|\omega_{ni}\|^2 \ell_k^2 c_k^2 1_{\{|\epsilon_i| + \ell_k c_k > \delta / \|\omega_{ni}\|\}} \right],$$

and both terms go to zero. Indeed, for the first term we have

$$\begin{aligned} & nE \left[ \|\omega_{ni}\|^2 E \left[ \epsilon_i^2 \{ |\epsilon_i| + c_k \ell_k > \delta \sqrt{n/\xi_k} \} | x_i \right] \right] \\ & \leq nE \left[ \|\omega_{ni}\|^2 \right] \cdot \sup_{x \in \mathcal{X}} E \left[ \epsilon_i^2 \{ |\epsilon_i| + c_k \ell_k > \delta \sqrt{n/\xi_k} \} | x_i = x \right] \\ & \leq \underline{\sigma}^{-2} o(1) = o(1) \end{aligned}$$

where we used (A.23), the uniform integrability in A.4 and  $\delta \sqrt{n/\xi_k} - c_k \ell_k \rightarrow \infty$ ; and for the second term

$$\begin{aligned} & nE \left[ \|\omega_{ni}\|^2 \ell_k^2 c_k^2 P \left[ |\epsilon_i| + c_k \ell_k > \delta \sqrt{n/\xi_k} | x_i \right] \right] \\ & \leq nE \left[ \|\omega_{ni}\|^2 \ell_k^2 c_k^2 \right] \cdot \sup_{x \in \mathcal{X}} P \left[ |\epsilon_i| + c_k \ell_k > \delta \sqrt{n/\xi_k} | x_i = x \right] \\ & \leq \underline{\sigma}^{-2} \ell_k^2 c_k^2 \cdot \frac{\bar{\sigma}^2}{[\delta \sqrt{n/\xi_k} - \ell_k c_k]^2} = o(1) \end{aligned}$$

where we used (A.23),  $\delta \sqrt{n/\xi_k} - c_k \ell_k \rightarrow \infty$  and  $c_k \ell_k = o(\delta \sqrt{n/\xi_k})$ . □

#### A.4. Proofs of Section 4.3.

*Proof of Lemma 2.* Decompose

$$\sqrt{n}\alpha(x)'(\hat{\beta} - \beta) = \alpha(x)'\mathbb{G}_n[p_i(\epsilon_i + r_i)] + \alpha(x)'[\hat{Q}^{-1} - I]\mathbb{G}_n[p_i(\epsilon_i + r_i)].$$

Step 1. Conditional on the data, let  $T := \{t = (t_1, \dots, t_n) \in \mathbb{R}^n : t_i = \alpha(x)'(\hat{Q}^{-1} - I)p_i\epsilon_i, x \in \mathcal{X}\}$ . Define the norm  $\|\cdot\|_{n,2}$  on  $\mathbb{R}^n$  by  $\|t\|_{n,2}^2 = n^{-1} \sum_{i=1}^n t_i^2$ . Letting  $\eta_1, \dots, \eta_n$  be independent Rademacher random variables independent of the data, we have by Dudley's inequality (Dudley, 1967)

$$E_\eta[\sup_{x \in \mathcal{X}} |\alpha(x)'(\hat{Q}^{-1} - I)\mathbb{G}_n[\eta_i p_i \epsilon_i]|] \leq C \int_0^\theta \sqrt{\log N(\varepsilon, T, \|\cdot\|_{n,2})} d\varepsilon,$$

where  $\theta := 2 \sup_{t \in T} \|t\|_{n,2} = 2 \sup_{x \in \mathcal{X}} \|\alpha(x)'(\hat{Q}^{-1} - I)p_i\epsilon_i\|_{L^2(\mathbb{P}_n)} \leq 2 \max_{1 \leq i \leq n} |\epsilon_i| \|\hat{Q}^{-1} - I\| \|\hat{Q}\|^{1/2}$ . Since for any  $x, x' \in \mathcal{X}$ ,

$$\begin{aligned} & \|\alpha(x)'(\hat{Q}^{-1} - I)p_i\epsilon_i - \alpha(x')'(\hat{Q}^{-1} - I)p_i\epsilon_i\|_{L^2(\mathbb{P}_n)} \\ & \leq \|\alpha(x) - \alpha(x')\| \cdot \|(\hat{Q}^{-1} - I)p_i\epsilon_i\|_{L^2(\mathbb{P}_n)} \\ & \leq L_{1k} \xi_k \max_{1 \leq i \leq n} |\epsilon_i| \|\hat{Q}^{-1} - I\| \|x - x'\| \\ & =: L'_{1k} \max_{1 \leq i \leq n} |\epsilon_i| \|\hat{Q}^{-1} - I\| \|x - x'\|, \end{aligned}$$

we have

$$N(\varepsilon, T, \|\cdot\|_{n,2}) \leq \left( \frac{CL'_{1k} \max_{1 \leq i \leq n} |\epsilon_i| \|\hat{Q}^{-1} - I\|}{\varepsilon} \right)^d.$$

Thus we have

$$\int_0^\theta \sqrt{\log N(\varepsilon, T, \|\cdot\|_{n,2})} d\varepsilon \leq \max_{1 \leq i \leq n} |\epsilon_i| \|\hat{Q}^{-1} - I\| \int_0^{2\|\hat{Q}\|^{1/2}} \sqrt{d \log(CL'_{1k}/\varepsilon)} d\varepsilon.$$

By A.5, we have  $E[\max_{1 \leq i \leq n} |\epsilon_i| \mid X] \lesssim_P n^{1/m}$ . Since  $\|\hat{Q}^{-1} - I\| \lesssim_P \sqrt{\xi_k^2 \log n/n}$ ,  $\|\hat{Q}\| \lesssim_P 1$  and  $\log L'_{1k} \lesssim \log n$ , we have

$$\begin{aligned} E[\sup_{x \in \mathcal{X}} |\alpha(x)'(\hat{Q}^{-1} - I)\mathbb{G}_n[p_i\epsilon_i]| \mid X] & \leq 2E[E_\eta[\sup_{x \in \mathcal{X}} |\alpha(x)'(\hat{Q}^{-1} - I)\mathbb{G}_n[\eta_i p_i \epsilon_i]| \mid X]] \\ & \lesssim_P n^{1/m} \sqrt{\frac{\xi_k^2 \log^2 n}{n}}, \end{aligned}$$

where the first inequality is due to the symmetrization inequality. Thus, we have

$$\sup_{x \in \mathcal{X}} |\alpha(x)'(\hat{Q}^{-1} - I)\mathbb{G}_n[p_i\epsilon_i]| \lesssim_P n^{1/m} \sqrt{\frac{\xi_k^2 \log^2 n}{n}}.$$

Step 2. Observe that

$$\sup_{x \in \mathcal{X}} |\alpha(x)'(\widehat{Q}^{-1} - I)\mathbb{G}_n[p_i r_i]| \leq \|\widehat{Q}^{-1} - I\| \sup_{\|\gamma\|=1} |\mathbb{G}_n[\gamma' p_i r_i]|.$$

We wish to bound  $\sup_{\|\gamma\|=1} |\mathbb{G}_n[\gamma' p_i r_i]|$ . Recall that  $E[p_i r_i] = 0$ . For any  $\|\gamma\| = 1$ ,

$$\begin{aligned} \left| \sum_{i=1}^n \gamma' p(x_i) r(x_i) \right| &= \left| \sum_{i=1}^n \sum_{j=1}^k \gamma_j p_j(x_i) r(x_i) \right| \\ &= \left| \sum_{j=1}^k \gamma_j \left( \sum_{i=1}^n p_j(x_i) r(x_i) \right) \right| \\ &\leq \sqrt{\sum_{j=1}^k \gamma_j^2} \sqrt{\sum_{j=1}^k \left( \sum_{i=1}^n p_j(x_i) r(x_i) \right)^2} \\ &= \sqrt{\sum_{j=1}^k \left( \sum_{i=1}^n p_j(x_i) r(x_i) \right)^2}. \end{aligned}$$

Taking expectation, we have

$$\begin{aligned} E\left[ \sup_{\|\gamma\|=1} |\mathbb{G}_n[\gamma' p_i r_i]| \right] &\leq \frac{1}{\sqrt{n}} E \left[ \sqrt{\sum_{j=1}^k \left( \sum_{i=1}^n p_j(x_i) r(x_i) \right)^2} \right] \\ &\leq \frac{1}{\sqrt{n}} \sqrt{\sum_{j=1}^k E \left[ \left( \sum_{i=1}^n p_j(x_i) r(x_i) \right)^2 \right]} \\ &\leq \ell_k c_k \sqrt{E[\|p(x_1)\|^2]} \wedge \xi_k c_k \leq \ell_k c_k \sqrt{k}. \end{aligned}$$

Thus, we have

$$\sup_{x \in \mathcal{X}} |\alpha(x)'(\widehat{Q}^{-1} - I)\mathbb{G}_n[p_i r_i]| \lesssim_P \sqrt{\frac{\xi_k^2 \log n}{n}} \ell_k c_k \sqrt{k}.$$

Steps 1 and 2 give the first linearization result.

Step 3. We wish to bound  $\sup_{x \in \mathcal{X}} |\alpha(x)' \mathbb{G}_n[p_i r_i]|$ . We use Theorem 12. Consider the class of functions

$$\mathcal{F} := \{\alpha(x)' p(\cdot) r(\cdot) : x \in \mathcal{X}\}.$$

Then,  $|\alpha(x)' p(\cdot) r(\cdot)| \leq \ell_k c_k \xi_k$  and for any  $x, \tilde{x} \in \mathcal{X}$ ,

$$|\alpha(x)' p(\cdot) r(\cdot) - \alpha(\tilde{x})' p(\cdot) r(\cdot)| \leq \ell_k c_k L_{1k} \xi_k \|x - \tilde{x}\|,$$

so that

$$\sup_Q N(\mathcal{F}, L^2(Q), \varepsilon \ell_k c_k \xi_k) \leq \left( \frac{CL_{1k}}{\varepsilon} \right)^d.$$

Thus, by Theorem 12, we have

$$E[\sup_{x \in \mathcal{X}} |\alpha(x)' \mathbb{G}_n[p_i r_i]|] \lesssim \ell_k c_k \sqrt{\log n} + \ell_k c_k \frac{\xi_k \log n}{\sqrt{n}} \lesssim \ell_k c_k \sqrt{\log n},$$

where we have used the fact that

$$\frac{\xi_k \log n}{\sqrt{n}} = \sqrt{\log n} \sqrt{\frac{\xi_k^2 \log n}{n}} = o(\sqrt{\log n}).$$

Therefore, we have

$$\sup_{x \in \mathcal{X}} |\alpha(x)' \mathbb{G}_n[p_i r_i]| \lesssim_P \sqrt{\log n} \ell_k c_k. \quad (\text{A.24})$$

This completes the proof.  $\square$

*Proof of Theorem 3.* We wish to bound  $\sup_{x \in \mathcal{X}} |\alpha(x)' \mathbb{G}_n[p_i \epsilon_i]|$ . To this end, we use Proposition 3. Consider the class of functions

$$\mathcal{G} := \{(\epsilon, x) \mapsto \epsilon \alpha(v)' p(x) : v \in \mathcal{X}\}.$$

Then,  $|\alpha(v)' p(x_i)| \leq \xi_k$ ,  $\text{var}(\alpha(v)' p(x_i)) = 1$  and for any  $v, \tilde{v} \in \mathcal{X}$ ,

$$|\epsilon \alpha(v)' p(x) - \epsilon \alpha(\tilde{v})' p(x)| \leq |\epsilon| L_{1k} \xi_k \|v - \tilde{v}\|.$$

Thus, taking  $G(\epsilon, x) := |\epsilon| \xi_k$ , we have

$$\sup_Q N(\mathcal{G}, L^2(Q), \varepsilon \|G\|_{L^2(Q)}) \leq \left( \frac{CL_{1k}}{\varepsilon} \right)^d.$$

Therefore, by Proposition 3, we have

$$E[\sup_{x \in \mathcal{X}} |\alpha(x)' \mathbb{G}_n[p_i \epsilon_i]|] \lesssim \sqrt{\log n} + \frac{\xi_k^{m/(m-2)} \log n}{\sqrt{n}} \lesssim \sqrt{\log n}, \quad (\text{A.25})$$

where we have used assumption A.5

$$\frac{\xi_k^{m/(m-2)} \log n}{\sqrt{n}} = \sqrt{\log n} \cdot \sqrt{\frac{\xi_k^{2m/(m-2)} \log n}{n}} \lesssim \sqrt{\log n}.$$

This completes the proof.  $\square$

*Proof of Theorem 4.* The proof follows similarly to that in Chernozhukov et al. (2009) and has two steps: in the first, we couple the estimator  $\sqrt{n}(\hat{\beta} - \beta)$  with the normal vector; in the second, we establish the strong approximation for the series estimate of the function.

STEP 1. We shall apply Yurinskii's coupling (see Theorem 10 in Pollard (2002)):

Let  $\zeta_1, \dots, \zeta_n$  be independent  $K$ -vectors with  $E[\zeta_i] = 0$  for each  $i$ , and  $\Delta := \sum_i E\|\zeta_i\|^3$  finite. Let  $S$  denote a copy of  $\zeta_1 + \dots + \zeta_n$  on a sufficiently rich probability space  $(\Omega, \mathcal{A}, P)$ . For each  $\delta > 0$  there exists a random vector  $T$  in this space with a  $N(0, \text{var}(S))$  distribution such that

$$P\{\|S - T\| > 3\delta\} \leq C_0 B \left(1 + \frac{|\log(1/B)|}{K}\right) \text{ where } B := \Delta K \delta^{-3},$$

for some universal constant  $C_0$ .

In order to apply the coupling, consider a copy of the first order approximation to our estimator on a suitably rich probability space

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \zeta_i, \quad \zeta_i = \Omega^{-1/2} p_i(\epsilon_i + r_i) \sim N(0, I_k),$$

Then since  $\text{maxeig}(\Omega^{-1/2})$  is bounded by an earlier argument,

$$\begin{aligned} E\|\zeta_i\|^3 &\lesssim E\|p_i(\epsilon_i + r_i)\|^3 \\ &\lesssim E[\|p_i\|^3(|\epsilon_i|^3 + |r_i|^3)] \\ &\lesssim E[\|p_i\|^3](1 + \ell_k^3 c_k^3) \\ &\lesssim E[\|p_i\|^2] \xi_k (1 + \ell_k^3 c_k^3) \\ &\lesssim k \xi_k (1 + \ell_k^3 c_k^3) \end{aligned}$$

where we used the assumption that  $E[|\epsilon_i|^3 | x_i]$  are uniformly bounded. Therefore, by Yurinskii's coupling, for each  $\delta > 0$

$$\begin{aligned} &P\left\{\left\|\frac{\sum_{i=1}^n \zeta_i}{\sqrt{n}} - \mathcal{N}_k\right\| \geq 3\delta a_n^{-1}\right\} \\ &\lesssim \frac{nk^2 \xi_k (1 + \ell_k^3 c_k^3)}{(\delta a_n^{-1} \sqrt{n})^3} \left\{1 + \frac{\log(k^2 \xi_k (1 + \ell_k^3 c_k^3))}{k}\right\} \\ &\lesssim \frac{a_n^3 k^2 \xi_k (1 + \ell_k^3 c_k^3)}{(\delta n^{1/2})} \left\{1 + \frac{\log n}{k}\right\} \rightarrow 0 \\ &\text{by } \frac{a_n^3 k^4 \xi_k^2 (1 + \ell_k^3 c_k^3)^2 \log^2 n}{n} \rightarrow 0, \end{aligned}$$



Finally by combining the preceding step with the assumption on the linearization error  $R_{1n}$ , we obtain for a copy of  $\sqrt{n}(\hat{\beta} - \beta)$  on a suitably rich probability space that obeys

$$\begin{aligned} \|\Omega^{-1/2}\sqrt{n}(\hat{\beta} - \beta) - \mathcal{N}_k\| &\leq \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \zeta_i - \mathcal{N}_k \right\| + \|\Omega^{-1/2}\sqrt{n}(\hat{\beta} - \beta) - \frac{1}{\sqrt{n}} \sum_{i=1}^n \zeta_i\| \\ &\leq o_P(a_n^{-1}) + R_{1n} = o_P(a_n^{-1}). \end{aligned}$$

This proves the first part of the theorem.

STEP 2. Using the result of Step 1 and that

$$\frac{\sqrt{np}(x)'(\hat{\beta} - \beta)}{\|s_n(x)\|} = \frac{\sqrt{n}s_n(x)'\Omega^{-1/2}(\hat{\beta} - \beta)}{\|s_n(x)\|}$$

we conclude that

$$|S_n(x)| := \left| \frac{\sqrt{n}s_n(x)'\Omega^{-1/2}(\hat{\beta} - \beta)}{\|s_n(x)\|} - \frac{s_n(x)'\mathcal{N}_k}{\|s_n(x)\|} \right|$$

satisfies

$$\sup_{x \in \mathcal{X}} |S_n(x)| \leq \left\| \sqrt{n}\Omega^{-1/2}(\hat{\beta} - \beta) - \mathcal{N}_k \right\| = o_P(a_n^{-1}), \quad (\text{A.26})$$

Finally,

$$\begin{aligned} &\sup_{x \in \mathcal{X}} \left| \frac{\sqrt{n}(\hat{g}(x) - g(x))}{\|s_n(x)\|} - \frac{s_n(x)'\mathcal{N}_k}{\|s_n(x)\|} \right| \\ &\leq \sup_{x \in \mathcal{X}} \left| \frac{\sqrt{n}(\hat{g}(x) - g(x))}{\|s_n(x)\|} - \frac{\sqrt{n}s_n(x)'\Omega^{-1/2}(\hat{\beta} - \beta)}{\|s_n(x)\|} \right| \\ &\quad + \sup_{x \in \mathcal{X}} \left| \frac{\sqrt{n}s_n(x)'\Omega^{-1/2}(\hat{\beta} - \beta)}{\|s_n(x)\|} - \frac{s_n(x)'\mathcal{N}_k}{\|s_n(x)\|} \right| \\ &= \sup_{x \in \mathcal{X}} |\sqrt{nr}(x)/\|s_n(x)\|| + \sup_{x \in \mathcal{X}} |S_n(x)| = o_P(a_n^{-1}) + o_P(a_n^{-1}), \end{aligned}$$

using the assumption on the approximation error  $r(x) = g(x) - p_n(x)'\beta$  and the bound (A.26).  $\square$

*Proof of Theorem 5.* Note that  $\hat{\beta}^b$  solves the least squares problem for the rescaled data  $\{(\sqrt{h_i}y_i, \sqrt{h_i}p_i) : 1 \leq i \leq n\}$ . The weight  $h_i$  is independent of  $(y_i, p_i)$ ,  $E[h_i] = 1$ ,  $E[h_i^2] = 1$ , and  $\max_{1 \leq i \leq n} h_i \lesssim_P \log n$ . That allows us to extend all results from  $\hat{\beta}$  to  $\hat{\beta}^b$  replacing  $\xi_k$  by  $\xi_k^b = \xi_k \log n$  to account for the larger envelope, and  $p_i^b = h_i p_i$ .

We apply Lemma 2 to the original problem and to the weighted problem by  $\{h_i\}$ . Then

$$\begin{aligned} \sqrt{n}(\hat{\beta}^b - \hat{\beta}) &= \sqrt{n}(\hat{\beta}^b - \beta) + \sqrt{n}(\beta - \hat{\beta}) \\ &= \mathbb{G}_n[(h_i - 1)p_i(\epsilon_i + r_i)] + r_n \end{aligned}$$

where  $\|r_n\| \lesssim_P \sqrt{\frac{\xi_k^2 \log^4 n}{n}} (n^{1/m} \sqrt{\log n} + \sqrt{k \log n} \ell_k c_k)$ .

Note also that the results continue to hold in  $P$ -probability if we replace  $P$  by  $P^*$ , since if a random variable  $B_n \lesssim_P 1$  then  $B_n \lesssim_{P^*} 1$ . Indeed, the first relation means that  $P(|B_n| > \ell_n) = o(1)$  for any  $\ell_n \rightarrow \infty$ , while the second means that  $P^*(|B_n| > \ell_n) = o_P(1)$  for any  $\ell_n \rightarrow \infty$ . But the second clearly follows from the first from the Markov inequality, observing that  $E[P^*(|B_n| > \ell_n)] = P(|B_n| > \ell_n) = o(1)$ .

The second part of the theorem follows similarly to Theorem 4 by applying Yurinskii coupling for the weighted process  $v_i = h_i - 1$ , where  $h_i \sim \exp(1)$  so that  $E[v_i^2] = 1$ ,  $E[|v_i|^3] \lesssim 1$  and  $E[\max_{1 \leq i \leq n} |v_i|] \lesssim \log n$ . Thus there is a Gaussian process  $G_n \sim N(0, I_k)$  such that

$$\left\| \frac{\Omega^{-1/2}}{\sqrt{n}} \sum_{i=1}^n (h_i - 1) p_i (\epsilon_i + r_i) - G_n \right\| \lesssim_P o(1/\log n).$$

Combining the result above with the first part of the theorem, the second part follows by the triangle inequality.  $\square$

*Proof of Theorem 6.* Under A.1 and A.2, we have that  $Q$  and  $\Sigma$  have eigenvalues bounded away from zero and from above uniformly in  $n$ , which implies that so does  $\Omega$ .

Under our growth conditions, the first result follows from the Markov inequality and Lemma 4 to establish  $E[\|\widehat{Q} - Q\|] \lesssim \|Q\| \sqrt{\xi_k^2 \log n / n} = o(1)$ .

To establish the second result we note that

$$\widehat{\Sigma} - \Sigma = \mathbb{E}_n[(\widehat{\epsilon}_i^2 - \{\epsilon_i + r_i\}^2) p_i p_i'] + \mathbb{E}_n[\{\epsilon_i + r_i\}^2 p_i p_i'] - \Sigma. \quad (\text{A.27})$$

The first term on the right hand side satisfies

$$\begin{aligned} & \|\mathbb{E}_n[(\widehat{\epsilon}_i^2 - \{\epsilon_i + r_i\}^2) p_i p_i']\| \leq \|\mathbb{E}_n[\{p_i'(\widehat{\beta} - \beta)\}^2 p_i p_i']\| + 2\|\mathbb{E}_n[(\epsilon_i + r_i) p_i'(\widehat{\beta} - \beta) p_i p_i']\| \\ & \lesssim \sup_{x \in \mathcal{X}} |p_i'(\widehat{\beta} - \beta)|^2 \|\mathbb{E}_n[p_i p_i']\| + \max_{i \leq n} \{|\epsilon_i| + |r_i|\} \sup_{x \in \mathcal{X}} |p_i'(\widehat{\beta} - \beta)| \cdot \|\mathbb{E}_n[p_i p_i']\| \rightarrow_P 0, \\ & \lesssim_P \|\widehat{Q}\| \frac{\xi_k^2 (\sqrt{\log n} + R_{1n} + R_{2n})^2}{n} + (v_n + \ell_k c_k) \|\widehat{Q}\| \frac{\xi_k (\log n + R_{1n} + R_{2n})}{\sqrt{n}} \rightarrow_P 0, \end{aligned}$$

since  $\sup_{x \in \mathcal{X}} |p_i'(\widehat{\beta} - \beta)|^2 \lesssim_P \xi_k^2 (\sqrt{\log n} + R_{1n} + R_{2n})^2 / n$  by Theorem 3,  $\|\mathbb{E}_n[p_i p_i']\| \lesssim_P 1$  by the first result,  $\max_{i \leq n} |r_i| \leq \ell_k c_k$ , and  $\max_{i \leq n} |\epsilon_i|^2 \lesssim_P v_n^2$  by Markov inequality.

To control the last terms in (A.27), note that

$$\begin{aligned}
 E[|\mathbb{E}_n[\{\epsilon_i + r_i\}^2 p_i p'_i] - \Sigma|] &\leq_{(1)} 2E E_\epsilon[|\mathbb{E}_n[\epsilon_i \{\epsilon_i + r_i\}^2 p_i p'_i]|] \\
 &\leq_{(2)} 2\sqrt{\frac{\log n}{n}} E[(|\mathbb{E}_n[\{\epsilon_i + r_i\}^4 \|p_i\|^2 p_i p'_i|])^{1/2}] \\
 &\leq_{(3)} 2\sqrt{\frac{\log n}{n}} \xi_k E[\max_{1 \leq i \leq n} |\epsilon_i + r_i| (|\mathbb{E}_n[\{\epsilon_i + r_i\}^2 p_i p'_i|])^{1/2}] \\
 &\leq_{(4)} 2\sqrt{\frac{\log n}{n}} \xi_k (E[\max_{1 \leq i \leq n} |\epsilon_i + r_i|^2])^{1/2} (E[|\mathbb{E}_n[\{\epsilon_i + r_i\}^2 p_i p'_i|])^{1/2},
 \end{aligned}$$

where (1) holds by Symmetrization Lemma (Lemma 2.3.6 van der Vaart and Wellner (1996)), (2) by Khinchin inequality, (3) by  $\max_{1 \leq i \leq n} \|p_i\| \leq \xi_k$ , and (4) by Cauchy-Scharwz.

Since that for any positive numbers,  $a \leq R(a+b)^{1/2}$  implies  $a \leq R^2 + R\sqrt{b}$ , the expression above using the triangle inequality yields

$$E[|\mathbb{E}_n[\{\epsilon_i + r_i\}^2 p_i p'_i] - \Sigma|] \leq \frac{4\xi_k^2 \log n}{n} (v_n^2 + \ell_k^2 c_k^2) + \left( \frac{4\xi_k^2 \log n}{n} \{v_n^2 + \ell_k^2 c_k^2\} \right)^{1/2} \|\Sigma\|^{1/2}.$$

The result follows from the Markov inequality.  $\square$

#### A.5. Proofs of Section 5.1.

*Proof of Theorem 7.* By Lemma 1 and P1

$$\begin{aligned}
 |\widehat{\theta}(w) - \theta(w)| &\leq |\ell(w)'(\widehat{\beta} - \beta)| + |r_n(w)| \\
 &\leq \frac{|\ell(w)' \mathbb{G}_n[p_i \epsilon_i]|}{\sqrt{n}} + \frac{\|\ell(w)\|(|R_{1n}| + |R_{2n}|)}{\sqrt{n}} + \|\ell(w)\| r_n(w) / \sqrt{n} \\
 &\leq \frac{|\ell(w)' \mathbb{G}_n[p_i \epsilon_i]|}{\sqrt{n}} + \frac{\xi_\theta(k, w)}{\sqrt{n}} [|R_{1n}| + |R_{2n}|] + o(\xi_\theta(k, w) / \sqrt{n})
 \end{aligned}$$

where the last inequality follows by  $\|\ell(w)\| \lesssim \xi_\theta(k, w)$  assumed in P2.

Next note that by Lemma 1

$$|R_{1n}| + |R_{2n}| \lesssim_P \sqrt{\frac{\xi_k^2 \log n}{n}} (1 + \sqrt{k \log n} \ell_k c_k) + \ell_k c_k = o(1).$$

Finally, since  $E[|\ell(w)' \mathbb{G}_n[p_i \epsilon_i]|^2] \lesssim \|\ell(w)\|^2 \|Q\| \lesssim \xi_\theta^2(k, w)$ , the result follows by applying the Chebyshev inequality to establish that  $|\ell(w)' \mathbb{G}_n[p_i \epsilon_i]| \lesssim_P \xi_\theta(k, w)$ .  $\square$

*Proof of Theorem 8.* First note that by A.1 and A.2,  $\Omega$  has eigenvalues bounded away from zero and from above. Moreover, under A.2, by Theorem 2 and P1

$$t_n(w) = \frac{\ell(w)'(\widehat{\beta} - \beta)}{\widehat{\sigma}_n(w)} + \frac{r_n(w)}{\widehat{\sigma}_n(w)} = \frac{\ell(w)' \Omega^{1/2} \mathcal{N}_k}{\sqrt{n} \widehat{\sigma}_n(w)} + o_P\left(\frac{\|\Omega^{1/2} \ell(w)\|}{\sqrt{n} \widehat{\sigma}_n(w)}\right) + o_P\left(\frac{\|\ell(w)\|}{\sqrt{n} \widehat{\sigma}_n(w)}\right).$$

To show that the last two terms are  $o_P(1)$ , note that by Theorem 6,  $\widehat{\sigma}_n(w) \gtrsim_P \|\Omega^{1/2}\ell(w)\|/\sqrt{n}$  since  $\widehat{\sigma}_n(w) = (1 + o_P(1))\sigma_n(w)$ .

Also because  $\widehat{\sigma}_n(w) = (1 + o_P(1))\sigma_n(w)$ , the first term satisfies

$$\frac{\ell(w)'\Omega^{1/2}\mathcal{N}_k}{\sqrt{n}\widehat{\sigma}_n(w)} \rightarrow_d N(0, 1).$$

□

### A.6. Proofs of Section 5.2.

*Proof of Theorem 9.* By the triangle inequality

$$\sup_{w \in \mathcal{I}} |\widehat{\theta}(w) - \theta(w)| \leq \sup_{w \in \mathcal{I}} |\ell(w)'(\widehat{\beta} - \beta)| + \sup_{w \in \mathcal{I}} |r_n(w)|$$

where the second term satisfies  $\sup_{w \in \mathcal{I}} |r_n(w)|/\|\ell(w)\| = o(n^{-1/2} \log^{-1} n)$  by condition U.1.

By Lemma 2, the first term is bounded uniformly over  $\mathcal{I}$  by

$$|\ell(w)'(\widehat{\beta} - \beta)| \lesssim_P \frac{|\ell(w)'\mathbb{G}_n[p_i(\epsilon_i + r_i)]|}{\sqrt{n}} + o_P(\xi_\theta(k, \mathcal{I})/[\sqrt{n} \log n]) \quad (\text{A.28})$$

since  $\|\ell(w)\| \leq \xi_\theta(k, \mathcal{I})$  by U.2 and the remainder term of the linear representation in Lemma 2 satisfies  $\sup_{w \in \mathcal{I}} \|r_n(w)\| = o_P(1/\log n)$ .

The result follows as the proof of Lemma 2 (A.24) and Theorem 3 (A.25) to establish

$$\sup_{w \in \mathcal{I}} \left| \frac{\ell(w)'}{\|\Omega^{1/2}\ell(w)\|} \mathbb{G}_n[p_i(\epsilon_i + r_i)] \right| \lesssim_P \sqrt{\log n} + \sqrt{\log n} \ell_k c_k.$$

□

*Proof of Theorem 10.* Under our conditions we have  $\|\widehat{\Omega} - \Omega\| = o_P(1/\sqrt{k} \log^{3/2} n)$  and  $\widehat{\sigma}_n(w) = (1 + o_P(1/\sqrt{k} \log^{3/2} n))\sigma_n(w)$  uniformly in  $w \in \mathcal{I}$ . Then the result follows from Theorem 4 to obtain the Gaussian approximation. □

*Proof of Theorem 11.* Let  $\varepsilon_n = 1/\log n$ , and  $\delta_n$  such that  $\delta_n \log^{1/2} n \rightarrow 0$ , and  $\delta_n/\varepsilon_n \rightarrow \infty$ .

Let  $\bar{t}_n^*(w) = \frac{\ell(w)'\Omega^{1/2}\mathcal{N}_k/\sqrt{n}}{\sigma_n(w)}$ ,  $w \in \mathcal{I}$ . Under our conditions we have  $\|\widehat{\Omega} - \Omega\| = o_P(1/\sqrt{k} \log^{3/2} n)$  and  $\widehat{\sigma}_n(w) = (1 + o_P(1/\sqrt{k} \log^{3/2} n))\sigma_n(w)$  uniformly in  $w \in \mathcal{I}$ . Then it follows that  $\|\bar{t}_n^*\|_{\mathcal{I}} = \sup_{w \in \mathcal{I}} |\bar{t}_n^*(w)|$ , which does not depend on the data  $\mathcal{D}_n$ , is such that

$$P(|\|\bar{t}_n^*\|_{\mathcal{I}} - \|t_n^*\|_{\mathcal{I}}| \leq \varepsilon_n) = 1 - o(1).$$

Now let  $k_n(1 - \alpha) := (1 - \alpha) - \text{quantile of } \|t_n^*\|_{\mathcal{I}}, \text{ conditional on } \mathcal{D}_n$ , and let  $\kappa_n(1 - \alpha) := (1 - \alpha) - \text{quantile of } \|\bar{t}_n^*\|_{\mathcal{I}}, \text{ conditional on } \mathcal{D}_n$ .

Then applying Lemma 6 to  $\|t_n^*\|_{\mathcal{I}}$  and  $\|t_n^*\|_{\mathcal{I}}$ , we get that for some  $\nu_n \searrow 0$

$$P[\kappa_n(p) \geq k_n(p - \nu_n) - \varepsilon_n \text{ and } k_n(p) \geq \kappa_n(p - \nu_n) - \varepsilon_n] = 1 - o(1).$$

Claim (1) now follows by noting that

$$\begin{aligned} P\{\|t_n\|_{\mathcal{I}} > k_n(1 - \alpha) + \delta_n\} &\leq P\{\|t_n\|_{\mathcal{I}} > \kappa_n(1 - \alpha - \nu_n) - \varepsilon_n + \delta_n\} + o(1) \\ &\leq P\{\|t_n^*\|_{\mathcal{I}} > \kappa_n(1 - \alpha - \nu_n) - 2\varepsilon_n + \delta_n\} + o(1) \\ &\leq P\{\|t_n^*\|_{\mathcal{I}} > k_n(1 - \alpha - 2\nu_n) - 3\varepsilon_n + \delta_n\} + o(1) \\ &\leq P\{\|t_n^*\|_{\mathcal{I}} > k_n(1 - \alpha - 2\nu_n)\} + o(1) \\ &= E_P[P\{\|t_n^*\|_{\mathcal{I}} > k_n(1 - \alpha - 2\nu_n) | \mathcal{D}_n\}] + o(1) \\ &\leq E_P[\alpha + 2\nu_n] + o(1) = \alpha + o(1). \end{aligned}$$

Claim (2) follows from the equivalence of the event  $\{\theta(w) \in [i(w), \ddot{i}(w)], \text{ for all } w \in \mathcal{I}\}$  and the event  $\{\|t_n\|_{\mathcal{I}} \leq c_n(1 - \alpha)\}$ .

To prove Claim (3) note that  $\hat{\sigma}_n(w) = (1 + o_P(1))\sigma_n(w)$  uniformly in  $w \in \mathcal{I}$  under our conditions by Theorem 6. Moreover,  $c_n(1 - \alpha) = k_n(1 - \alpha)(1 + o_P(1))$  because  $1/k_n(1 - \alpha) \lesssim_P 1$  and  $\delta_n \rightarrow 0$ . Combining these relations the result follows.

Claim (4) follows from Claim (1) and from the following lower bound.

By Lemma 6, we get that for some  $\nu_n \searrow 0$

$$P[\kappa_n(p + \nu_n) + \varepsilon_n \geq k_n(p) \text{ and } k_n(p + \nu_n) + \varepsilon_n \geq \kappa_n(p)] = 1 - o(1).$$

Then

$$\begin{aligned} P\{\|t_n\|_{\mathcal{I}} \geq k_n(1 - \alpha) + \delta_n\} &\geq P\{\|t_n\|_{\mathcal{I}} \geq \kappa_n(1 - \alpha + \nu_n) + \varepsilon_n + \delta_n\} - o(1) \\ &\geq P\{\|t_n^*\|_{\mathcal{I}} \geq \kappa_n(1 - \alpha + \nu_n) + 2\varepsilon_n + \delta_n\} - o(1) \\ &\geq P\{\|t_n^*\|_{\mathcal{I}} \geq k_n(1 - \alpha + 2\nu_n) + 3\varepsilon_n + \delta_n\} - o(1) \\ &\geq P\{\|t_n^*\|_{\mathcal{I}} \geq k_n(1 - \alpha + 2\nu_n) + 2\delta_n\} - o(1) \\ &\geq E[P\{\|t_n^*\|_{\mathcal{I}} \geq k_n(1 - \alpha + 2\nu_n) + 2\delta_n | \mathcal{D}_n\}] - o(1) \\ &= \alpha - 2\nu_n - o(1) = \alpha - o(1), \end{aligned}$$

where we used the anti-concentration property in the last step.  $\square$

### A.7. Proofs of Section 6.

*Proof of Lemma 4.* Using the Symmetrization Lemma (Lemma 2.3.6 van der Vaart and Wellner (1996)) and the Khinchin inequality, bound

$$\Delta := E\|\widehat{Q} - Q\| \leq 2EE_\varepsilon\|\mathbb{E}_n[\varepsilon_i Q_i]\| \leq \sqrt{\frac{\log n}{n}} E\|(\mathbb{E}_n Q_i^2)^{1/2}\|$$

Since

$$E\|(\mathbb{E}_n Q_i^2)^{1/2}\| = E\|(\mathbb{E}_n Q_i^2)\|^{1/2} \leq \left[ME\|\mathbb{E}_n Q_i\|\right]^{1/2}$$

and

$$\|\mathbb{E}_n Q_i\| \leq \Delta + \|Q\|,$$

one has

$$\Delta \leq \sqrt{\frac{M \log n}{n}} [\Delta + \|Q\|]^{1/2},$$

solving which for  $\Delta$  gives the result stated in the lemma.  $\square$

*Proof of Proposition 3.* For a  $\tau > 0$  specified later, define  $\epsilon_i^- := \epsilon_i I(|\epsilon_i| \leq \tau) - E[\epsilon_i I(|\epsilon_i| \leq \tau) | X_i]$  and  $\epsilon_i^+ := \epsilon_i I(|\epsilon_i| > \tau) - E[\epsilon_i I(|\epsilon_i| > \tau) | X_i]$ . Since  $E[\epsilon_i | X_i] = 0$ ,  $\epsilon_i = \epsilon_i^- + \epsilon_i^+$ . Invoke the decomposition

$$\sum_{i=1}^n \epsilon_i f(X_i) = \sum_{i=1}^n \epsilon_i^- f(X_i) + \sum_{i=1}^n \epsilon_i^+ f(X_i).$$

We apply Theorem 12 to the first term. Noting that  $\text{var}(\epsilon_i^- f(X_i)) \leq \sup_x E[(\epsilon_i^-)^2 | X_i = x] E[f(X_i)^2] \leq \sup_x E[\epsilon_i^2 | X_i = x] = \sigma^2$  and  $\epsilon_i^- f(X_i) \leq 2\tau b$ , we have

$$E \left[ \left\| \sum_{i=1}^n \epsilon_i^- f(X_i) \right\|_{\mathcal{F}} \right] \leq C \left[ \sqrt{n\sigma^2 V \log(Ab)} + V\tau b \log(Ab) \right].$$

On the other hand, applying Theorem 2.14.1 of van der Vaart and Wellner (1996) to the second term, we obtain

$$E \left[ \left\| \sum_{i=1}^n \epsilon_i^+ f(X_i) \right\|_{\mathcal{F}} \right] \leq C \sqrt{nb} \sqrt{E[|\epsilon_1^+|^2]} \int_0^1 \sqrt{V \log(A/\epsilon)} d\epsilon. \quad (\text{A.29})$$

By assumption,

$$E[|\epsilon_1^+|^2] \leq E[\epsilon_1^2 I(|\epsilon_1| > \tau)] \leq \tau^{-m+2} E[|\epsilon_1|^m],$$

by which we have

$$(A.29) \leq C \sqrt{E[|\epsilon_1|^m]} b \tau^{-m/2+1} \sqrt{nV \log(A)}.$$

Taking  $\tau = b^{2/(m-2)}$ , we obtain the desired inequality.  $\square$

### A.8. Additional technical results.

**Lemma 5.** *Let  $Z$  be a random vector in  $\mathbb{R}^k$ ,  $M$  be a  $k \times k$  matrix and  $\Gamma \subset \mathbb{R}^k \setminus \{0\}$ . Then we have that*

$$\sup_{\gamma \in \Gamma} E \left[ \left| \frac{\gamma'}{\|\gamma\|} M Z \right|^m \right] \leq \|M\|^m \sup_{\|a\|=1} E [|a'Z|^m].$$

*Proof of Lemma 5.* Let  $\bar{\gamma}$  achieve the supremum on the left hand side and set  $\bar{a} = \bar{\gamma}/\|\bar{\gamma}\|$ . Then we have

$$\begin{aligned} E [|\bar{a}' M Z|^m] &= E [|(M'\bar{a})' Z|^m] = \|M'\bar{a}\|^m E \left[ \left| \frac{(M'\bar{a})'}{\|M'\bar{a}\|} Z \right|^m \right] \\ &\leq \|M'\bar{a}\|^m \|\bar{a}\|^k E \left[ \left| \frac{(M'\bar{a})'}{\|M'\bar{a}\|} Z \right|^m \right] \\ &\leq \|M\|^m \sup_{\|a\|=1} E [|a'Z|^m] \end{aligned}$$

since  $\|\bar{a}\| = 1$  and  $M'\bar{a}/\|M'\bar{a}\| = 1$ .  $\square$

**Lemma 6** (Closeness in Probability Implies Closeness of Conditional Quantiles). *Let  $X_n$  and  $Y_n$  be random variables and  $\mathcal{D}_n$  be a random vector. Let  $F_{X_n}(x|\mathcal{D}_n)$  and  $F_{Y_n}(x|\mathcal{D}_n)$  denote the conditional distribution functions, and  $F_{X_n}^{-1}(p|\mathcal{D}_n)$  and  $F_{Y_n}^{-1}(p|\mathcal{D}_n)$  denote the corresponding conditional quantile functions. If  $|X_n - Y_n| = o_P(\varepsilon)$ , then for some  $\nu_n \searrow 0$  with probability converging to one*

$$F_{X_n}^{-1}(p|\mathcal{D}_n) \leq F_{Y_n}^{-1}(p + \nu_n|\mathcal{D}_n) + \varepsilon \text{ and } F_{Y_n}^{-1}(p|\mathcal{D}_n) \leq F_{X_n}^{-1}(p + \nu_n|\mathcal{D}_n) + \varepsilon, \forall p \in (\nu_n, 1 - \nu_n).$$

*Proof of Lemma 6.* We have that for some  $\nu_n \searrow 0$ ,  $P\{|X_n - Y_n| > \varepsilon\} = o(\nu_n)$ . This implies that  $P[P\{|X_n - Y_n| > \varepsilon|\mathcal{D}_n\} \leq \nu_n] \rightarrow 1$ , i.e. there is a set  $\Omega_n$  such that  $P(\Omega_n) \rightarrow 1$  and  $P\{|X_n - Y_n| > \varepsilon|\mathcal{D}_n\} \leq \nu_n$  for all  $\mathcal{D}_n \in \Omega_n$ . So, for all  $\mathcal{D}_n \in \Omega_n$

$$F_{X_n}(x|\mathcal{D}_n) \geq F_{Y_n+\varepsilon}(x|\mathcal{D}_n) - \nu_n \text{ and } F_{Y_n}(x|\mathcal{D}_n) \geq F_{X_n+\varepsilon}(x|\mathcal{D}_n) - \nu_n, \forall x \in \mathbb{R},$$

which implies the inequality stated in the lemma, by definition of the conditional quantile function and equivariance of quantiles to location shifts.  $\square$

### REFERENCES

- Andrews, D.W.K. (1991). Asymptotic normality of series estimators for nonparametric and semiparametric models. *Econometrica* **59** 307-345.
- Angrist, J., Chernozhukov, V. and Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the U.S. wage structure. *Econometrica* **74** 539-563.

- Burman, P., Chen, K.W. (1989). Nonparametric estimation of a regression function. *Annals of Statistics* **17** 1567-1596.
- Chen, X. (2006). Large sample sieve estimation of semi-nonparametric models. In: Heckman, J.J., Leamer, E. (Eds.), *Handbook of Econometrics*, vol. 6B. Elsevier (Chapter 76).
- Chernozhukov, C., Fernández-Val, I. and Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica* **78** 1093-1125.
- Chernozhukov, V., Lee, S. and Rosen, A. (2009). Intersection bounds: estimation and inference. arXiv:0907.3503.
- Cohen, A., Daubechies, I., and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal* **1** 54-81.
- DeVore, R.A. and Lorentz, G.G. (1993). *Constructive Approximation*. Springer.
- Dudley, R. M. (1967). The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *J. Functional Analysis*, **1**, 290-330.
- Eastwood, B.J., Gallant, A.R. (1991). Adaptive rules for seminonparametric estimation that achieve asymptotic normality. *Econometric Theory* **7** 307-340.
- Gallant, A.R., Souza, G. (1991). On the asymptotic normality of Fourier flexible functional form estimates. *Journal of Econometrics* **50** 329-353.
- Giné, E. and Koltchinskii, V. (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.* **34** 1143-1216.
- Guédon O. and Rudelson, M. (2007).  $L_p$ -moments of random vectors via majorizing measures. *Advances in Mathematics* **208** 798-823.
- Horowitz, J.L. (2009). *Semiparametric and Nonparametric Methods in Econometrics*. Springer.
- Huang, J.Z. (2003a). Asymptotics for polynomial spline regression under weak conditions. *Statist. Probab. Lett.* **65** 207-216.
- Huang, J.Z. (2003b). Local asymptotics for polynomial spline regression. *Ann. Statist.* **31** 1600-1635.
- Lust-Picard L. and Pisier, G. (1991). Non-commutative Khintchine and Paley inequalities. *Arkiv för Matematik* **29** 241-260.
- Mallat, S. (2009). *A Wavelet Tour of Signal Processing*. Third Edition. Academic Press.
- Massart, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.* **28** 863-884.
- Newey, W.K. (1997). Convergence rates and asymptotic normality for series estimators. *J. Econometrics* **79** 147-168.



- Pollard, D. (2002). A User's Guide to Measure Theoretic Probability. Cambridge Series in Statistics and Probabilistic Mathematics.
- Rudelson, M. (1999). Random vectors in the isotropic position. *Journal of Functional Analysis* **164**, 1, 60-72.
- Stone, C.J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. With discussion by Andreas Buja and Trevor Hastie and a rejoinder by the author. *Ann. Statist.* **22** 118-184.
- Talagrand, M. (1996a). Majorizing measures: the generic chaining. *Ann. Probab.* **24** 1049–1103.
- Talagrand, M. (1996b). New concentration inequalities in product spaces. *Invent. Math.* **126** 505–563.
- Tsybakov, A.B. (2003). *Introduction to Nonparametric Estimation*. Springer.
- Wasserman, L. (2006). *All of nonparametric statistics*. Springer.
- van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press.
- van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.